

Kristina Š. Despot
Ana Ostroški Anić
Ivana Brač (Eds.)

Lexicography and Semantics



Proceedings of the
XXI EURALEX International Congress

8–12 October 2024
Cavtat, Croatia



LEXICOGRAPHY AND SEMANTICS

Nakladnik:	Institut za hrvatski jezik
Za nakladnika:	Željko Jozić
Knjiga:	Lexicography and Semantics, Proceedings of the XXI EURALEX International Congress
Urednice:	Kristina Štrkalj Despot Ana Ostroški Anić Ivana Brač
Tehničko uređenje i računalni slog:	Elena Vrbanić
Oblikovanje naslovnice:	Elena Vrbanić, with the help of OpenAI
Tisak:	



© This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License, 2024.

ISBN 978-953-7967-77-2

Kristina Š. Despot, Ana Ostroški Anić, Ivana Brač (Eds.)

Lexicography and Semantics

Proceedings of the
XXI EURALEX International Congress

8–12 October 2024
Cavtat, Croatia

TABLE OF CONTENTS

FOREWORD	13

PART I: CONFERENCE OVERVIEW

Acknowledgements	17
Main Sponsors.....	17
Sponsors.....	18
Programme Committee.....	18
Local Organising Committee.....	18
Scientific Committee.....	18
Overview of Keynotes and Workshops.....	21
Keynotes.....	21
Keynotes of the Workshop Figurative Language and Large Language Models.....	21
Workshops.....	21

PART II: PROCEEDINGS

Chapter I. Lexicography and Semantics	25

Tony Veale

You Talk Funny! Someday Me Talk Funny Too!	
– On Learning to See the Humorous Side of Familiar Words.....	27

Valeria Caruso, Lucia di Pace

Words for Choosing Food in <i>ALMA – Multimedia Atlas of Bio/Cultural Food</i>	41

Janet DeCesaris, Mercè Lorente Casafont

Old Words, New Terms – Semantic Broadening and	
Narrowing in the Vocabulary of the Circular Economy.....	55

Maucha Gamonal, Adriana Pagano, Tiago Torrent, Ely Matos, Arthur Lorenzi

Automated Semantic Frame Annotation –	
An Exploratory Study in the Health Domain.....	67

Ellert Thor Johannsson, Thordis Ulfarsdottir

The Role of Semantic Fields in Contemporary Icelandic Dictionaries.....	81

Robert Krovetz

Morpho-Semantics and Dictionary Entries.....	93

Haniva Yunita Leo

Do Indonesian Speakers Feel ‘pain’? NSM and Corpus-Based	
Approach to the Cross-Linguistic Concept of ‘pain’ in Bahasa Indonesia.....	109

Michael Nguyen, Peter Juel Henriksen

The Locum Hyphen – A Formal Approach to the Lexicalization	
of Multiword Expressions With Rich Internal Semantics.....	121

Pär Nilsson

Figurative, Transferred or Extended use? The Use of Semantic Labels in the
First Edition and the Revised Version of the Swedish Academy Dictionary.....131

Sanni Nimb, Ida Flörke, Sussi Olsen, Bolette S. Pedersen, Nathalie C. H. Sørensen

COR.SEM, a New Formal Semantic Lexicon for Danish.....141

Andrej Perdih, Kozma Ahačič, Nataša Jakop, Nina Ledinek, Špela Petric Žižić

Semantic Information on the Franček Educational Language Portal for Slovenian.....155

Emma Sköldberg, Shafqat Mumtaz Virk, Pauline Sander,**Simon Hengchen and Dominik Schlechtweg**

Revealing Semantic Variation in Swedish Using Computational Models of
Semantic Proximity – Results From Lexicographical Experiments.....169

Petra Storjohann

Synonyms in Contrast – A Dynamic and Descriptive
Resource for New Semantic Equivalents.....183

Lars Trap-Jensen, Henrik Lorentzen

Back to Basics – Meaning Description for Human Users and for Computers.....191

Ene Vainik, Geda Paulsen, Heete Sahkai, Jelena Kallas,**Arvi Tavast, Arvi Tavast, Kristina Koppel**

From a Dictionary to a Constructicon – Putting the Basics on the Map.....209

Chapter II. Lexicography and Language Technologies.....217**Igor Boguslavsky**

Lexical Resources for a Semantic Parser of Russian
– Argument Structure of Ordinal Adjectives.....219

Nataliia Cheilytko, Ruprecht von Waldenfels

Word Embeddings for Detecting Lexical Semantic Change in Ukrainian.....231

Lian Chen, Wenjun Sun, Flora Badin

Innovation in Phraseomatics – DiCoP Project and DiCoP-Text Corpus
for the Enrichment of Language Models and Automatic Translation.....243

Enikő Héja, Kata Gábor, László Simon, Veronika Lipp

Graph-based Detection of Hungarian Adjectival Meaning
Structures via Monolingual Static Embeddings.....251

Geraint Paul Rees, Isabel Gibert

A Textbook or ChatGPT – Which Helps Novice
Programmers Most with Unknown Terms?.....265

Bálint Sass

The “Dependency Tree Fragments” Model for Querying a Constructicon.....275

Dragana Špica, Benedikt Perak

Enhancing Japanese Lexical Networks Using Large Language

Models – Extracting Synonyms and Antonyms with GPT-4o.....283

Chapter III. Dictionary Writing Systems and Lexicographic Tools.....303**Michaela Denisová, Gilles-Maurice de Schryver, Pavel Rychlý**

The Automatic Determination of Translation Equivalents in

Lexicography: What Works and What Doesn't?.....305

Stephanie Evert, Christine Ganslmayer, Christian Rink

Multi-Level Analysis as a Systematic Approach to

Evaluating the Quality of AI-Generated Dictionary Entries.....317

Carolina Flinz, Daniel Henkel, Valeria Zotti, Sabrina Ballestracci

A Multilingual Parallel Corpus for the Lexical Information

System LBC – Recent Progress and Future Perspectives.....335

Peter Juel Henriksen

Make Each Morph Count – A New Approach to

Computational Lexicography for Text Processing.....349

František Kovařík, Vojtěch Kovář, Marek Blahuš

On Rapid Annotation of Czech Headwords

– Analysing the First Tasks of Czech Dictionary Express.....357

Irene Renau, Rogelio Nazar, Daniel Mora Melanchthon

Towards the Automatic Generation of a Pattern-Based Dictionary of Spanish Verbs.....367

Chapter IV. Reports on Lexicographical and Lexicological Projects.....383**Juris Baldunčiks, Silga Sviķe**Pages of Latvian Historical Slang Dictionary: *dzeršana* ('drinking').....385**María Auxiliadora Barrios Rodríguez**

Diretes, a Spanish Monolingual Dictionary Based on Lexical-Semantic Relations.....393

Luke Omoyemi Akinremi, María José Domínguez Vázquez

The Erasmus Mundus Joint Masters in Lexicography

– EMJM-EMLex – New Developments and Goals.....407

Dwayne Ellul

Maltese Lexicography – A Historical Context and the Current State.....413

Ivana Filipović Petrović, Kristina Kocijan

Creating the Dataset of Croatian Verbal Idioms – Automatic

Identification in a Corpus and Lexicographic Implementation.....429

Louise Holmer, Ann Lillieström, Emma Sköldberg, Jonatan Uppström Time to Say Goodbye Revisited – On the Exclusion of Headwords from the Swedish Academy Glossary (SAOL).....	443
Kathryn Hudson The Benefits of Bio(lexicography) – A Topical Approach to Lexicographic Practice.....	453
Boris Kern Considering Word Formation in Compiling Dictionaries.....	463
Anas Fahad Khan, Ana Salgado, Isuri Anuradha, Rute Costa, Chamila Liyanage, John P. McCrae, Atul K. Ojha, Priya Rani and Francesca Frontini Cultural Heritage and Multilingual Understanding through lexiCal Archives (CHAMUÇA) – Portuguese Borrowings in Contemporary Asian Languages.....	475
Veronika Kolářová, Jiří Mirovský Looking for Sense in Nonsense – Valency of Negative Forms of Nouns and Adjectives in the NomVallex Lexicon.....	485
Kusujiro Miyoshi John Pickering's Reference Materials for His <i>Vocabulary</i> (1816) – Transcending the Bounds of Dictionaries.....	497
Pär Nilsson Report on the Revision of the Swedish Academy Dictionary – and the Search for “Old Neologisms”.....	507
Sanni Nimb, Nathalie C. H. Sørensen, Jonas Jensen Making Danish Thesaurus Data Available to Researchers – The WebDDB project.....	523
Christian-Emil Smith Ore, Oddrun Grønvik The Spoken Word as Represented in Norsk Ordbok.....	531
Petya Osenova, Kiril Simov All About Words! An Integrated Dictionaries Portal for Bulgarian.....	547
Vanja Štefanec, Krešimir Šojat, Matea Filko CroDeriv – Search and Visualization Interface.....	557
Chapter V. Bi- and Multilingual Lexicography.....	563
Vladimír Benko, Zuzana Kříhová, Boris Lehečka, Darina Vystrčilová Persian to Czech Dictionary – A Traditional Dictionary in the Era of AI?.....	565
Cormac Breathnach, Pádraig Ó Mianáin Making a Molehill out of a Mountain: Technical and Editorial Considerations in Producing the Concise English-Irish Dictionary (2020).....	581
Elina Chadjiapapa, Zoe Gavriilidou Helix – A Bilingual Illustrated Dictionary for Greek Heritage Learners.....	597

Rajna Dragićević, Yury Makarov, Daria Ryzhova, Yulia Shapich, Ekaterina Yakushkina A New Serbian-Russian Dictionary.....	607
Mariusz Piotr Kamiński The Contribution of Bilingualized Entries and Vocabulary Knowledge to the Learner's Success in Sentence Completion – The Case of <i>jump</i> Verbs.....	615
Evelina Kirsakmene False Friends in General Bilingual Dictionaries (English and French into Latvian).....	623
Daria Lazić Lexicographic Treatment of Vocabulary Related to Age – The Example of Croatian and Danish.....	633
Irina Lobzhanidze, Rusudan Gersamia, Nino Tsulaia Compiling a Bilingual Megrelian-English Online Dictionary – Preserving Endangered Kartvelian Languages.....	647
Chenlu Yu A German-Chinese e-Dictionary of Manufacturing Technology in Automotive Industry – Entry Design.....	661
Chapter VI. Specialized Lexicography, Terminology, and Terminography.....	669
Andrea Abel, Natascia Ralli Gender in Electronic Dictionaries and Terminology Databases – State of the Art and Future Directions.....	671
Ieda Maria Alves, Beatriz Curti-Contessoto, Ana Maria Ribeiro de Jesus Challenges of Creating a Medical Dictionary for a Low Literacy Audience in Brazil – Focusing on Politically Marked Terms Related to the COVID-19 Pandemic.....	689
Lynne Bowker Eponyms, EDI and Terminology Planning in the Biological Sciences.....	699
Theresa Kruse, Ulrich Heid, Barbara Schmidt-Thieme Mathematics Students as Lexicographers – Learning Domain Concepts and Their Relations by Designing Dictionary Articles and Concept Maps.....	711
David Lindemann Teaching Terminology Through Wikibase and Wikidata.....	719
Bruno Nahod Can We Substitute Field Experts with Customized Large Language Model in Processing Specialized Languages? – A Case Study.....	725

Chapter VII. Dictionary (in) Use.....739

Margit Langemets, Lydia Risberg, Kristel Algreve

To Dream or Not to Dream About 'Correct' Meanings?

– Insights into the User Experience Survey.....741

Tinatin Margalitadze, Katalin P. Márkus

Cross-Border Collaboration in Teaching Dictionary Skills.....761

Barbora Štěpánková, Lucie Poláková, Jana Šindlerová, Michal Novák

What Can Dictionaries Tell Us About Pragmatic Markers

– Building the Lexicon of Epistemic and Evidential Markers in Czech.....769

Chapter VIII. Historical and Dialect Lexicography.....783

David Lindemann, Mikel Alonso

Linking Historical Corpus Data and Annotations Using Wikibase.....785

Ivana Lovrić Jović, Martina Kramarić

The Dubrovnik Idiom Through Time – Crafting a Diachronic Dictionary.....791

Magdalena Majdak

Defining Meanings in Historical Dictionaries – The Case of the

Electronic Dictionary of the 17th- and 18th-Century Polish.....805

Martina Waclawičová

Dialect Dictionary and Lexicalization of Dialect Phenomena.....819

Leonardo Zilio, Besim Kabashi

Using Neural Machine Translation for Normalising Historical Documents.....827



EURALEX XXI
CROATIA 2024

FOREWORD

This book contains papers presented at the XXI EURALEX International Congress, Cavtat, Croatia, 8–12 October 2024.

The main topic of the XXI EURALEX International Congress was *Lexicography and Semantics*.

Lexicography and semantics, particularly lexical semantics, share a deep connection, as the accurate representation of meaning is a fundamental aspect of both disciplines. The intersection of lexicography and semantics is a fertile ground for exploring how words convey meaning and how meaning is structured in language.

Over the past century and a half, various significant theoretical approaches to understanding word meaning have left their mark on the field of lexicography. Notably, the influence of semantic theories on lexicography and the compilation of digital lexical resources has become particularly prominent with the rise of cognitive semantics, and more specifically, the advent of frame semantics, which has played a pivotal role in enhancing the way lexicographers analyse and describe word meanings, resulting in more comprehensive lexical resources.

These resources include traditional dictionaries, thesauri, and linguistic corpora, but also extend to more specialized databases, ontologies, word embeddings, and semantic networks. This remarkable range of lexical resources has had a profound impact on the field of semantics, stimulating new lines of research and enriching our understanding of how language works. These resources have found applications beyond these two linguistic fields: They are used in fields like information retrieval, sentiment analysis, machine translation, and Large Language Models development, demonstrating the practical relevance of the synergy between lexicography and semantics, which is very well represented at this conference.

EURALEX 2024 fostered valuable discussions, and promoted collaboration within the fields of lexicography and semantics, with a particular focus on cognitive approaches and their applications.

Papers, posters, presentations of new projects, and software demonstrations were invited on the topic *Lexicography and Semantics* (theoretical and applied perspectives) as well as on all lexicographically relevant topics including, but not limited to, the following fields:

- Monolingual, bilingual, and multilingual lexicography, including commercial dictionary publishing
- The dictionary-making process
- The use of dictionaries, and the needs of the user
- Lexicography and language technologies: the use of computational linguistics, natural language processing and computer science technologies and language resources in lexicography
- Terminology, terminography, and specialized forms of lexicography
- Lexicography of lesser-used and under-researched languages
- Phraseology and collocations

- Lexicography and etymology
- Lexicological issues of lexicographical relevance
- Reports on lexicographical and lexicological projects
- Academic lexicography, historical lexicography, scholarly lexicography
- Theoretical linguistic research, especially those branches most directly related to the interests and needs of compilers of dictionaries and/or computer lexicons, including word meaning, syntax, word combinations (phraseology), morphology, phonology, etc.

All submissions underwent a double-blind peer review process by at least two members of the Scientific Committee, whose support we deeply appreciate (see page 19). Decisions to accept or reject submissions for presentation at the congress and full papers for publication in the conference proceedings were based on the average review scores, and in many cases, further evaluation by members of the Programme Committee, to whom we are immensely grateful (see page 14). We extend our special gratitude to the EURALEX Board members who supported us as part of the Programme Committee: **Annette Klosa-Kückelhaus**, **Iztok Kosem**, **Robert Lew**, and **Philipp Stöckle**. Without the expertise and dedication of all colleagues serving on the Scientific and Programme Committees, we would not have been able to maintain the high academic standards of EURALEX congress presentations and proceedings. Thank you!

Our plenary speakers were invited to explore various aspects of the congress theme, *Lexicography and Semantics*. However, given the high relevance and timeliness of Large Language Models, it is no surprise that this topic resonates throughout all the keynotes and the entire conference program. We are deeply grateful to the keynote speakers for their timely, relevant, thought-provoking, and engaging talks.

In his plenary talk *Lexical Semantics, Lexicography and LLMs*, **Dirk Geeraerts** explores the evolving role of Large Language Models (LLMs) in lexical studies from both lexicographic and theoretical linguistic perspectives. From a theoretical perspective, LLMs currently constitute the epitome of distributional semantics, and distributional semantics is eminently suited as a methodological basis for usage-based cognitive semantics, allowing for a convergence of major theoretical trends in lexical semantics. From the lexicographical perspective, Geeraerts draws attention to the too often ignored process through which lexicography not only gave a major descriptive impetus to the development of corpus linguistics, but also specifically contributed to an essential step in the emergence of computational methods for corpus research: LLMs are a tool with lexicographic roots, at least to some extent. But, given that the tool has grown well beyond its original format, how does that affect its relationship to lexicography?

Tony Veale's talk *You Talk Funny! One Day Me Talk Funny Too* explores the intriguing challenge of teaching Large Language Models (LLMs) to understand and generate humour. While AI has made significant strides in various areas of human intelligence, humour remains a uniquely human trait that machines struggle to grasp. The talk emphasizes the importance of humour in AI, both for practical applications and in pushing the boundaries of human-machine interaction. It outlines current research efforts to equip LLMs with a “sense of humour” and highlights the complexities, such as navigating social norms and ensuring value alignment, given that humour often tests the limits of acceptability. The ultimate goal is to develop AI that can not only comprehend human jokes but also create its own, adding a new dimension to AI's capabilities.

In her talk *Case Studies in the Successes and Limits of Frame Semantics in Practical Lexicography*, **Kory Stamper** examines how frame semantics has influenced modern lexicography, particularly in the U.S., where traditional methods have long dominated. Frame semantics offers lexicographers a broader, more user-centered approach to defining meanings, helping them overcome unconscious biases. However, recent technological and social shifts, such as the spread of misinformation, the rise of AI-generated text, and the changing use of online dictionaries, present challenges in further integrating frame semantics. The talk highlights case studies showcasing both the successes and limitations of this approach in the evolving field of lexicography.

Tiago Torrent talks about *Possible Futures for Semantic Lexical Resources in the Age of Artificial Intelligence* exploring the future of combining semantic lexical resources with Large Language Models (LLMs) in language technology. While LLMs have made significant strides in NLP, recent research shows their limitations, suggesting that hybrid approaches integrating LLMs with language resources may offer more potential. Tiago highlights possible research directions, including the application of the FrameNet model in developing tools to identify regions prone to gender-based violence and expanding into the field of multimodal NLP. These examples illustrate how blending AI with semantic resources could drive future advancements in language technology.

Lana Hudeček offers us valuable insight into the perspective of a lexicographer and the intricate process of dictionary making. The presentation on the *Croatian Web Dictionary – Mrežnik* highlights the development and evolution of this project, which began in 2017 and is now funded by the EU's NextGeneration program. Mrežnik is a free, online, corpus-based, monolingual, and normative e-dictionary of the Croatian language, consisting of modules for native speakers, schoolchildren, and non-native speakers. Utilizing tools like Sketch Engine and TLex, it integrates contemporary e-lexicography practices and computational linguistics methods. Mrežnik aims to be dynamic and continually updated, integrating features like gamification and recorded pronunciation. The project addresses the lack of an e-dictionary in Croatia and focuses on user-centered design, ensuring accessibility and relevance for various audiences.

I would like to extend special thanks to the keynote speakers of the *Figurative Language and Large Language Models* workshop: **Eve Sweetser** and **Alexander Ziem**.

Eve Sweetser, who has long been both an academic and personal role model, as well as a dear friend, discusses the current status and future directions of the Berkeley MetaNet project. The new MetaNet Wiki is under construction, and will (fairly soon) become accessible, including databases of French, English, and Spanish metaphors relating to (1) COVID and the pandemic; (2) cancer; and (3) climate change. This project has all of the scholarly community's shared concerns with the methods of metaphor identification and analysis: the talk therefore recaps the MetaNet procedure for searching, identifying, and verifying data. The particular focus of the talk is on the comparative aspects of the project. How can we use the wealth of documentation of English metaphors as a resource, without allowing our knowledge about English to "colonialize" our other databases?

Alexander Ziem's talk focuses on (a) examining the role of conceptual metaphors (CM) within a construction, (b) explaining how CMs—understood as frames with conventionalized mappings between a source and target frame—have been conceptually integrated into the construction, and (c) detailing the empirical methods used to identify and implement CMs within the system.

This book contains the full papers from keynotes, talks, posters, and software demonstrations presented at the XXI EURALEX International Congress. A total of sixty-seven full papers were accepted for publication. An alphabetical index at the end of this volume lists all authors and their details. The papers are thematically divided into the following chapters:

Chapter I. Lexicography and Semantics

Chapter II. Lexicography and Language Technologies

Chapter III. Dictionary Writing Systems and Lexicographic Tools

Chapter IV. Reports on Lexicographical and Lexicological Projects

Chapter V. Specialized Lexicography, Terminology, and Terminography

Chapter VI. Bi- and Multilingual Lexicography

Chapter VII. Dictionary (in) Use

Chapter VIII. Historical Lexicography.

The congress was organized by the Institute for the Croatian Language (ICL) in Zagreb, Croatia. Our sincere thanks go to all our colleagues at the ICL who supported the organization of the congress and the publication of both the Book of Abstracts and the Proceedings. We are especially grateful to **Željko Jozić** for his unwavering support, not only as the director of the Institute but also as a member of the Local Organizing Committee.

We would also like to extend our heartfelt thanks to all the **sponsors** (see page 17, 18), whose financial support made EURALEX 2024 possible. Without their generosity, this congress would not have taken place.

As Chair of the XXI EURALEX Organizing Committee, I would like to express my deep gratitude to all members of our Local Organizing Committee (see page 19). I would also like to extend my heartfelt thanks to **Annette Klosa-Kückelhaus**, who has been invaluable to us both as EURALEX President and as the exceptional organizer of EURALEX 2022. Her help and support have made this journey far smoother, and without her, it would have been significantly more challenging. Finally, a huge thank you goes to my co-editors, **Ana Ostroški Anić** and **Ivana Brač**, whose tireless efforts were invaluable not only in bringing this publication to life but also in ensuring the success of the entire conference.

Kristina Š. Despot
Chair of EURALEX 2024
September 2024

PART I: CONFERENCE OVERVIEW

Acknowledgements

We are deeply grateful to all those who generously supported the XXI EURALEX International Congress financially. Your contributions have been invaluable, and we sincerely appreciate your commitment to making this event a success.

Main Sponsors



Sponsors



Republic
of Croatia
Ministry
of Culture
and Media
Republika
Hrvatska
Ministarstvo
kulture
i medija

We extend our heartfelt thanks to everyone who dedicated their time and expertise to reviewing the submissions and papers. Your invaluable contributions have greatly enriched the quality of our conference, and we are truly appreciative of your efforts.

We wish to express our deepest gratitude to the members of the Local Organizing Committee for their remarkable dedication and hard work in making this event a reality. We truly appreciate all the time and energy you've invested in creating a memorable and smooth experience for everyone.

Programme Committee

Annette Klosa-Kückelhaus
Iztok Kosem
Robert Lew
Philipp Stöckle
Ivana Brač
Milica Mihaljević
Ana Ostroški Anić
Kristina Š. Despot

Local Organising Committee

Kristina Š. Despot
Ana Ostroški Anić
Ivana Brač
Željko Jozić
Ivana Filipović Petrović
Ivana Lovrić Jović
Maja Matijević
Martina Pavić
Siniša Runjaić
Lobel Filipić

Scientific Committee

Andrea Abel
Arleta Adamska-Salaciak
Špela Arhar Holdt
Petra Bago
Martina Bajčić
Vincent Balnat
Vuk-Tadija Barbarić
Verginica Barbu Mititelu
Slobodan Beliga
Vladimír Benko
Goranka Blagus Bartolec
Tomislava Bošnjak Botica
Petar Božović
Ivana Brač
Maja Bratanić
Úna Bhreathnach

Thomas Burch
Valeria Caruso
Paul Cook
Rute Costa
Gilles-Maurice de Schryver
Janet Decesaris
Idalete Maria Silva Dias
María José Domínguez Vázquez
Anna Dziemianko
Pamela Faber
Ilse Feinauer
Ivana Filipović Petrović
Edward Finegan
Carolina Flinz
Thierry Fontenelle

Judit Freixa
Polona Gantar
Radovan Garabík
Zoe Gavriilidou
Alexander Geyken
Voula Giouli
Oddrun Grønvik
Volker Harm
Kris Heylen
Lana Hudeček
John Humbley
Max Ionov
Dubravka Ivšić Majić
Miloš Jakubiček
Jelena Kallas
Virna Karlić

Boris Kern	Milica Mihaljević	Ranka Stanković
Ilan Kernerman	Fabio Mollica	Egon W. Stemle
Fahad Khan	Carolin Müller-Spitzer	Frieda Steurs
Annette Klosa-Kückelhaus	Hilary Nesi	Philipp Stöckle
Svetla Koeva	Henrik Nilsson	Kristina Štrkalj Despot
Veronika Kolářová	Pär Nilsson	Giovanni Tallarico
Kristina Koppel	Vincent Ooi	Elsabe Taljard
Iztok Kosem	Christian-Emil Ore	Pius ten Hacken
Simon Krek	Petya Osenova	Carole Tiberius
Margit Langemets	Julia Ostanina-Olszewska	Lars Trap-Jensen
Barbara Lewandowska-	Ana Ostroški Anić	Anna Vacalopoulou
Tomaszczyk	Gokhan Ozkan	Carlos Valcárcel Riveiro
Robert Lew	Benedikt Perak	Tony Veale
Marie-Claude L'Homme	Sanja Perić Gavrančić	Federica Vezzani
Anja Lobenstein-Reichmann	Ralf Plate	Craig Volker
Irina Lobzhanidze	Boris Pritchard	Sabine Wahl
Henrik Lorentzen	Ida Raffaelli	Ana Werkmann Horvat
Magdalena Majdak	Natascia Ralli	Geoffrey Williams
Tinatin Margalidze	Geraint Paul Rees	Sascha Wolfer
Saša Marijanović	Irene Renau	Alexander Ziem
Ivana Matas Ivanković	Ana Salgado	Tanara Zingano Kuhn
John P. McCrae	Thomas Schmidt	Mojca Žagar Karer
Peter Meyer	Hindrik Sijens	Slavko Žitnik

Overview of Keynotes and Workshops

Keynotes

Geeraerts, Dirk	Lexicography and Theories of Lexical Semantics (And the Other Way Round)
Hudeček, Lana	Croatian Web Dictionary – MREŽNIK
Stamper, Kory	Case Studies in the Successes and Limits of Frame Semantics in Practical Lexicography
Torrent, Tiago	Possible Futures for Semantic Lexical Resources in the Age of Artificial Intelligence
Veale, Tony	You Talk Funny! One Day Me Talk Funny Too. Investigating the Capacity of Large Language Models for Humour

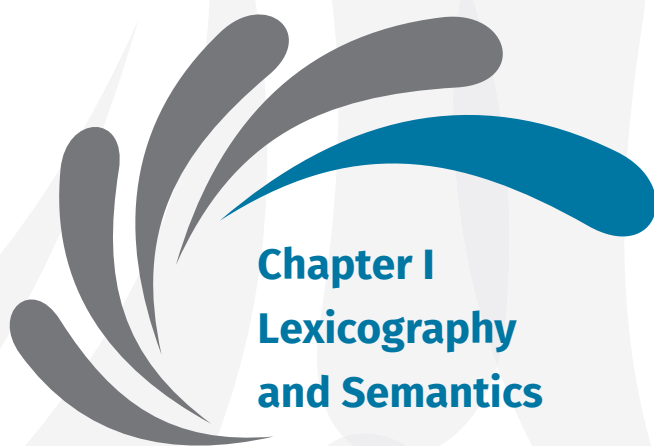
Keynotes of the Workshop *Figurative Language and Large Language Models*

Sweetser, Eve	METANET – Its Current Status and Future Directions
Ziem, Alexander	Conceptual Metaphors in the German FrameNet Construction: From Conception to Implementation

Workshops

Despot, Kristina Š. and Ana Ostroški Anić	Figurative Language and Large Language Models
Krek, Simon	Large Language Models and Lexicography
Reese, Gerraint Paul, Blanca Arias-Badia, Elisenda Bernal, and Sergi Torner	Lexicography and Accessibility

PART II: PROCEEDINGS



Chapter I
Lexicography
and Semantics

Tony Veale

YOU TALK FUNNY!

SOMEDAY ME TALK FUNNY TOO!

On Learning to See the Humorous Side of Familiar Words

Abstract Verbal humour brings a playful flexibility to our sober notions of meaning, truth and the mental lexicon. But before we can knowingly subvert this sober world view, we must learn the rules and conventions that define it. Traditionally, we have done this by building models that make explicit claims about words and meanings, but a new AI paradigm, large language models (LLMs), allows our models to tacitly learn much more about lexical semantics and pragmatics than anything we could hope to encode in a symbolic model. LLMs build context-specific encodings of words that are unique to, yet still generalizable from, their every attested use in a corpus. Humour, and the ability to generate novel jokes, stretch this ability to contextualize word meanings, just as they have severely tested our existing symbolic approaches. In this paper we explore whether LLMs like GPT-3.5 or GPT-4, which underpin the application ChatGPT, can “get” the intents of jokes that use familiar words in non-obvious ways, or whether, indeed, they can craft meaningful new jokes of their own.

Keywords humour; jokes; large language models; irony; sincerity

1. Introduction

1.1 Thinking Deep, Talking Funny

Whenever we set out to be humorous with language, we rarely use different words; rather, we use familiar words differently. It is not the words that differ, as such, but the rules of the language games in which we use them. Different families of joke will vary in the games they play, from narrative jokes to one-liners to puns, but in each case the teller and listener play by different rules. As tellers know where conformity will lead, they can see the conclusions to which an audience will leap. Yet the teller also knows those conclusions to be flawed, and has another, less obvious destination in mind for their audience. Like metaphors, jokes forge new links between familiar ideas (Pollio, 1996), and turn failed predictions into joyful realizations. Perhaps one must be able to feel joy and surprise to truly “get” a joke (Coulson & Kutas, 2001), but laughter is not the only way of showing one’s understanding. While we do not expect our AI systems, such as large language models (LLMs), to genuinely laugh at jokes, we do expect them to show insight into why, and how well, a joke works, and to be stable and discerning in these judgments. This is just one aspect, albeit a very important one, of what it means for a person or an LLM to have a sense of humour.

LLM-based agents such as ChatGPT already exhibit an amiable if facile wit of their own, and show a skilled ear for mimicry that lets them capture the cadences and attitudes of famous comics. Their auto-regressive generation of continuations to a given prompt makes them especially adept at using the “yes, and ...” principle of improv comedy to wring laughs from silly premises. But jokes of the short, snappy variety that are crafted to be retold in many different contexts require an LLM to do more than talk like a comedian. The words and attitudes must be appropriate to a joke, but the text must actually work like a joke too, with a logical mechanism that snaps shut on an audience like a mouse in a trap (Attardo et al., 1991; Attardo et al., 2002). Our focus in this paper is on jokes with a one-line setup and a one-line punchline, of the kind studied in Mihalcea & Strapparava (2010), that give LLMs little room to hide behind verbosity. This is the compressed format of the typical joke that speakers casually share and eagerly retell regardless of the context.

1.2 Collecting Funny: Gathering a Dataset of Jokes

Past work on joke analysis has prized this format for its ability to do more with less, whether as pithy one-liners, or tweets barbed with irony and sarcasm (Reyes et al., 2013; Veale, 2021), or satirical headlines from parody news sites like *The Onion* (West et al., 2019), or the captions adorning *The New Yorker’s* cartoons (Shahaf et al., 2015). For a rich source of jokes with popularity ratings, we turn to the *r/Jokes* community on Reddit, for which a dataset of over 100,000 jokes is available. These vary widely in quality, *upvote* counts (each is a thumbs-up from a user), joke type, and length, although all jokes have a header and a body that, for the short jokes we seek, correspond to the setup and punchline. We filter for jokes in which the header and body each comprise a single utterance, and only keep those whose header/setup is a simple question and whose body/punchline is a declarative answer. This focus on riddle-like jokes of the “*What do ...*” or “*Did you hear ...*” variety ensures that our dataset comprises many archetypal jokes and excludes observations, comments, anecdotes and other assorted pseudo-jokes that one finds in this sub-reddit. We also exclude the long tail of candidate jokes that have earned fewer than 3 upvotes.

We filter this much-reduced set to remove any that are vulgar, profane, racist, misogynist or homophobic, using a commercial moderation API from OpenAI. This allows us to rate jokes for offensiveness on multiple dimensions, and we exclude any that it flags as problematic. Curiously, upvote counts are no predictor of whether a joke will be flagged. For jokes with 5 or more upvotes, 10.2% are flagged; of those with 10 or more, 10.5% are flagged; of those with 50 or more, 11.6% are flagged; and of those with 100 or more, 12.2% are flagged. Whatever drives upvotes for jokes on Reddit, it is neither offensiveness nor its absence. We also filter duplicated jokes, and close variants, of which there are many. Two jokes are treated as variants if the cosine similarity of their vector encodings is greater than 0.9; we discard the variant with the fewest upvotes. The final, filtered dataset comprises 6,246 short question-answer jokes, of which the most popular (with 30,499 upvotes) is: “How do you milk sheep? With iPhone accessories.” As a rule, setups are longer than punchlines, which are conversely snappier than setups.

The mean setup length is 46 characters ($\mu=46.17$, $\sigma=14.61$) and the mean punchline length is 27 characters ($\mu=27.13$, $\sigma=17.1$). Only in 15% of cases is the punchline longer than its setup. Notwithstanding the occasional exception, what matters most in a punchline is pointed concision.

2. Appreciating and Understanding Jokes

2.1 Rating Jokes: The Raw and the Cooked

Joke generation is a language game that few people can play professionally, but joke appreciation is one we can all play. It is likewise easier for a system like ChatGPT to explain good jokes than to generate them itself. ChatGPT often demurs to rate humorousness directly, citing the subjectivity of enjoyment, so we instead prompt its LLM, GPT-3.5-Turbo, to rate the sophistication (which it seemingly deems more objective) of all 6,246 jokes in our dataset, on a scale of 0 to 100. We employ two variants of the basic prompt “*Rate the sophistication of this joke on a of 0 to 100: <joke>.*” The first, which asks for “*just a number*”, returns *raw ratings*, and so the LLM expends no extra tokens to justify its scores. The second remedies this lack by asking for a rating that is “*based on your understanding of the joke.*” We dub these *cooked ratings*, as the LLM expends considerably more tokens to justify its score.

Raw ratings are normally distributed about a mean of $\mu=57.7$ ($\sigma=24.7$), while cooked ratings have a mean of just $\mu=43.03$ ($\sigma=21.7$). The latter’s considered judgments are consistently lower than the former’s rapid ones by an average of 15 points. These two approaches reflect the System I and System II distinction of dual processing theory (Evans, 2003; Kahneman, 2011), with cooked ratings providing a System II corrective to the hasty judgments of System I. We observe that both sets of ratings are moderately correlated (Pearson’s $r = .53$). To ensure these judgments are stable over time, we later obtain a new set of ratings of each kind for each joke. We observe a good correlation between different sets of raw ratings ($r = .76$) and a strong correlation between sets of cooked ratings ($r = .82$). It seems we can expect cooked ratings to be the most stable, based as they are on reasoned judgments.

Cooked judgments also use the rating scale in a more human-like fashion, and show a marked preference for multiples of 10 (85%) over 5 (15%). These judgments effectively use a ten-point scale, opting for half-points for occasional nuance. Raw judgments show a preference for multiples of 5 (55%) over 10 (44%) and effectively use a twenty-point scale. To test ChatGPT’s discernment on real jokes, we create a set of 1000 incoherent jokes by randomly swapping setups and punchlines among jokes. ChatGPT’s cooked and raw ratings for this set show more agreement than for real jokes: the mean raw rating $\mu=33.8$ ($\sigma=26.2$) is just a few points higher than the mean cooked rating $\mu=30.3$ ($\sigma=18.9$). Cooked ratings for incoherent jokes are also highly correlated across runs at different times ($r = 0.81$), while the correlation between cooked and raw ratings remains consistent ($r = 0.53$). While incoherent jokes are rated lower on average, the difference is not significant enough (about one standard deviation) to suggest the LLM can discern real jokes from failed attempts.

Góes et al. (2023) report a weak positive correlation between the human ratings of jokes and those of GPT-4 when the LLM is instructed with a variety of different system messages. We also observe the effect on ratings of different system messages. When ChatGPT is instructed that “You are a professional comedian who hates puns”, its mean raw rating drops to 4.77. Conversely, when instructed that it “loves puns”, its mean raw rating rises to 91.44. These findings reflect the prevalence of wordplay in the *r/jokes* data. When ChatGPT is instructed to prize dark humour, its mean raw rating is 66 ($\sigma=18.7$), again reflecting the tenor of the Reddit dataset.

3. Generating Jokes

3.1 Talk Like A Comedian

Like many humans, ChatGPT rarely excels when it is put on the spot and asked to tell a joke apropos of nothing. Although it exhibits an easy wit and an ear for idiom and cliché, Jentzsch & Kersting (2023) observe that ChatGPT mostly – about 90% of the time – dips onto a small store of 25 or so famous funnies when it is prompted to tell a joke. We get more variety when we give it a topic, or better yet, a setup for it to complete, but even then it occasionally pulls an old chestnut from its bag.

As when rating jokes, we can ask ChatGPT to approach generation from the fast and slow perspectives of dual processing theory (Evans 2003). To invoke System I, we ask that a joke, or a punchline, is generated directly, without the expenditure of intermediate tokens and “compute.” To invoke System II, we ask that it approach the problem logically (Kahneman, 2011), and work through a chain of intermediate steps and output tokens to produce its results. Prompt engineering with exemplars elicits the best results for a System I approach. We can, for instance, instruct the LLM to respond to setups with an apt punchline, and provide it with an example:

- (1) Setup: Why do politicians take laxatives?
Punchline: So they can speak more fluently.

After priming the pump, we now present our own setup for the LLM to complete:

- (2) Setup: Why did the politician have an affair?
Punchline: To prove they could break promises on a personal level too.

To go beyond one- or few-shot priming, we can fine-tune the LLM on a large set of jokes. When we fine-tune GPT3.5-Turbo for 1 epoch on our dataset of 6,246 jokes, presented as *setup:* and *punchline:* pairs as shown above, we obtain a model that is more robust in its humour and more attuned to a joke’s cadences, as in this output:

- (3) Setup: Why did the politician have an affair?
Punchline: Because the debates were getting dull.

Fine-tuning on this scale with this kind of data has unintended consequences. We find, for instance, that ChatGPT is more likely to generate ungrammatical outputs

(as in “She wasn’t anybodies (sic) puppet.”) and vulgar and offensive ones (as in “She said to him, ‘vote for me and I’ll give you a BJ”). While our dataset has been carefully filtered, a joking attitude can free the *id* of an LLM from the fragile *super-ego* that additional value-alignment training has imposed upon it (Freud, 1905).

In Chain-of-Thought (CoT) prompting (Wei et al., 2022) we give an LLM a clear sequence of pseudo-code-like steps to follow. When programming at this high-level, the keenest insights come from professional comedy-writers who teach others how to systematically write jokes. Dean (2000) and Toplyn (2014) each propose a step-by-step process for writing new jokes. Dean’s process starts from a setup, chains thru a list of its connected topics, and concludes with the re-interpretation of one of these “connectors.” Toplyn’s process starts with an inspiring stimulus, such as a headline, and identifies two phrasal handles for the ideas that will be juxtaposed by a joke. This mirrors the theoretical view in Raskin (1984), Attardo (1994) and Attardo et al. (2002) that jokes involve a script opposition or a frame conflict (Coulson 2001) that motivates a sudden shift from one to the other. Toplyn (2022) has translated his professional intuitions into an LLM-based joke generator named *WitScript*.

In this vein (see Winters, 2023), ChatGPT can be given a WitScript-like script for composing a joke about a topic X, so that it pursues a specific chain of thought:

You are a professional joke writer for a comedian. To write a joke about a topic X, follow these steps:

1. *Identify three offbeat associations of X that we all know.*
2. *Identify a surprising link between each association in step 1 and X.*
3. *Turn the links in step 2 into joke setups, without revealing the surprise.*
4. *Turn the surprising aspect of each link in step 3 into a snappy punchline.*
5. *Select the least predictable punchline from step 4.*
6. *Assemble the joke by pairing the setup and punchline.*

Because its LLM has been fine-tuned to follow complex instructions (Ouyang et al., 2020), ChatGPT can bind a given topic to X and execute each step in turn. Notice how the key steps require insight (1, 2), imagination (2), skill (3), and appreciation (4, 5). Thus, when its topic is *awkward first dates*, ChatGPT suggests three frames, *uncomfortable silences*, *overthinking greetings*, and *selecting a date spot*, and three secondary perspectives, *crickets chirping*, *mental gymnastics* and *solving a Rubik’s cube*. It opts for the second of each, greetings & gymnastics, to produce this joke:

(4) Setup: Ever notice how overthinking greetings on first dates is like preparing for a diplomatic summit?

Punchline: Turning a handshake into a mental gymnastics routine should be an Olympic sport.

The end-result is facile and overly verbose, in both setup and punchline; each is more than two standard deviations longer than the mean lengths in our fine-tuning dataset. Nonetheless, it is a coherent joke with a keen sense of human social rituals.

Toplyn's *WitScript* is given a recent stimulus from which to shape a topical joke, but in a null context, we need a more timeless topic to feed to, or inspire, the LLM. Our dataset contains a multitude, from lying politicians and condescending hipsters to faithless husbands and nagging wives. We use these to instruct ChatGPT on the concerns of classic jokes, and ask it to suggest N more in the same vein. In response, its LLM shows a firm grasp of who and what our jokes typically take aim at, from corrupt CEOs to nutty conspiracy theorists, and its topic suggestions include many familiar targets, such as car salesmen, in-laws, bad drivers, new technology and first dates. We will use this capacity to drive the generation of jokes in the next section.

4. Evaluating Approaches to Joke Generation

4.1 When is a Joke not a Joke?

Humour is not an objective property of a text. Rather, the status of *joke* is ascribed to a text by its audience, in a process of interpretation that identifies what Raskin et al. (2009) call the *humour potential* of the text. The methods we consider here, from direct, System I approaches to CoT-based System II approaches, produce texts of varying shapes and even more variable humour potentials. Some LLM outputs *look* like jokes but come up short on humour potential, because their punchlines are incoherent or because, as in the case of this non-joke “What do you call a person who speaks multiple languages? A *polyglot*”, they present a too-sincere response to a disingenuous setup. Some work as jokes, if perhaps only weakly, while others may repeat an existing joke from the LLM's pre-training or fine-tuning data. To see how the approaches stack up, and how ChatGPT's self-rating of jokes stacks up against human judgments, we generate 1000 jokes using each of the following approaches:

1. *One-shot priming*: A fixed exemplar of a *setup*: prompt with a *punchline*: response (the politicians & laxatives joke) is presented to ChatGPT. A setup is then generated from a topic, and given to the LLM with the prefix *setup*:
2. *One-shot priming with length sampling*: As in 1, but the LLM's response is discarded and resampled if a punchline is longer than its setup, or if the setup or punchline lengths are more than one standard deviation longer than the corresponding means in our dataset.
3. *Fine-tuned one-shot priming with sampling*: As in 2, but the LLM (GPT3.5-Turbo) is first fine-tuned on the 6,246 short jokes in our filtered dataset.
4. *RAG with sampling*: A setup is generated from a topic, and RAG (retrieval-augmented generation) (Lewis et al., 2020) retrieves the two most similar jokes in our dataset, as ranked by cosine similarity of setups. The LLM is primed with these two retrieved jokes before it is given the new setup with the prefix *setup*: The LLM's responses are again filtered for length, as in 2.

5. *Chain-of-Thought (CoT)*: ChatGPT is prompted with a given topic and the 6-step CoT script in section 3.2. Of the 3 joke candidates produced by the script, only the LLM's preferred joke is retained.

GPT3.5-Turbo is used to test each of the approaches. In each case, the LLM is first primed with the instruction "You are a professional joke writer for a comedian" Topics are elicited from the LLM in batches of 20 with this prompt: "A classic joke will focus on a topic such as cheating husbands, crooked lawyers, lying politicians, and so on. Suggest 20 new topics for more jokes." In all but approach 5, the LLM is then prompted to generate a setup for each topic, and this setup is presented to the LLM to elicit a punchline. Table 1 reports the mean concision and balance for the outputs of each approach. A joke is deemed *concise* if its setup and punchline are within one standard deviation of the mean lengths in our dataset, and is *balanced* when its punchline is shorter than its setup. To show the effect of fine-tuning and RAG on these factors, concision and balance are reported for approaches 3 and 4 (in parentheses) before the LLM's outputs are sampled for length and concision.

Table 1 also reports the mean raw and cooked ratings that GPT3.5-Turbo self-assesses for each batch of 1000 outputs. Notice that interventions into how jokes are generated, such as post-hoc sampling for concision and balance (approach 2), fine-tuning (3) and priming with RAG (4) yield lower returns as far as the LLM itself is concerned. The least concise approaches tend to produce the best scores in Table 1.

Table 1: Percentage of outputs from each LLM-based approach that are concise and balanced; and the LLM's mean raw and cooked self-assessments of those outputs with standard deviations

Approach	Concision	Balance	Raw $\mu(\sigma)$	Cooked $\mu(\sigma)$
1. One-shot priming	20.3%	66.2%	72.34 (9.9)	61.24 (14.08)
2. Length sampling	100%	100%	66.65 (12.69)	51.97 (18.1)
3. Fine-tuning	(35.5%)	(88.7%)	60.15 (17.53)	46.02 (20.84)
4. RAG (two-shot)	(46.7%)	(58.8%)	66.85 (12.21)	54.06 (16.92)
5. CoT 6-step script	22.4%	72.6%	72.56 (10.69)	64.95 (12.14)

However, LLM self-assessments are not always reliable indicators of joke quality. Non-jokes and incoherent jokes often rate highly, with LLMs bringing a generosity of interpretation to dubious jokes that human judges withhold. When we manually assess each joke candidate from each approach, we discover a different story. Most candidates convey a flip or sardonic attitude, but do not resemble the humorous earworms that we might reuse in new settings, or that audiences will want to retell to others. Nonetheless, when labelling joke candidates by hand, we apply a low bar for acceptability: if a text's jocular intention is clear, and it exhibits a coherently realized humorous potential in the vein of Raskin (1984), we label it as a joke.

Table 2: Percentage of manually-rated outputs that are novel (new) vs. remembered (old) jokes, and percentages of LLM outputs that are deemed too incoherent or too sincere to work as jokes

Approach	New Jokes	Old Jokes	Incoherent	Sincere
1. One-shot priming	3.9%	1.5%	2.2%	4.1%
2. Length sampling	19.3%	11.3%	2%	2.8%
3. Fine-tuning	11.1%	10.6%	27.3%	19.4%
4. RAG (two-shot)	19.6%	3.1%	4.8%	10.3%
5. CoT 6-step script	6.5%	0%	5%	4.6%

Acceptable jokes are divided into two categories in Table 2: novel jokes (new) that are invented by the LLM itself, and pre-existing (old) jokes that it remembers and retells. To identify the latter, we search not just our dataset, but the wider internet. For instance, web search suggests that “What did the puzzle app say to the Sudoku app? I’ve got you all figured out.” (approach 2, sampling for concision and balance) is a new joke, while “How do you stop a politician from drowning? Shoot him before he hits the water” (approach 3, fine-tuning) is one that the LLM merely recalls.

The results in Table 2 contrast sharply with the LLM’s self-assessments in Table 1. In our manual assessment, RAG-based intervention (approach 4) yields the best returns: 1 in 5 of its outputs works as a new joke, and its invention-to-recall ratio is high. Priming the pump with retrieved exemplars actually seems to deter recall and regurgitation, since an LLM will be guided by these exemplars but will not want to replicate them. Fine-tuning (approach 3) also yields strong returns, but it may not repay its investment in time, resources and higher API costs. Approach 2, post-hoc sampling of the LLM’s outputs for concision and balance, yields the best return for the least effort. Filtering candidates on their superficial resemblance to jokes yields the greatest number of acceptable jokes overall (30.6%), but one in three of these (11.3%) are old familiars that were very likely seen during the LLM’s pre-training. CoT (approach 5) and priming with a fixed exemplar (approach 1) yield the least number of acceptable jokes, due in large part to their untamed verbosity, but this also ensures they are unlikely to include snappy, pre-existing jokes in their outputs.

We reject many joke candidates due to their lack of coherence or jocular intent. A punchline will seem incoherent if it fails to build on the setup, or if it refers to a conceit not grounded in the setup, or if it seems to belong to another joke, as in this output from approach 4 (RAG): “What do you call a telemarketer who won’t stop talking? A telephone-tortoise.” A punchline lacks clear jocular intent when it offers a sincere, perhaps literal, response to the setup, as in this output from approach 3 (fine-tuning): “Did you hear about the actor who forgot his lines on stage? It was the last time he had a part in a play.” Table 2 reports the percentage of candidates from each approach that we deemed too incoherent or too sincere to be jokes.

The greatest percentage of incoherent outputs (27.3%) is produced by fine-tuning with sampling for concision and balance (approach 3). We suspect that fine-tuning on a large set of jokes disrupts the LLM’s already well-tuned sensibilities, and the pressure to produce snappy responses leads it to sacrifice coherence for concision. This approach also has the most overly-sincere outputs (19.4%). There are other forms of humour, and other spurs to laughter, than the prototypical joke (Provine, 2000; Scott et al., 2014). Many sincere responses have a seed of humour within them, but fail to rise to the level of a joke, such as this output of fine-tuning : “How did the parents handle their child’s first breakup? They got through it together.” A sincere response may be true but mildly ironic, as in this other output of approach 3: “How did the meeting to discuss work-life balance go? It dragged on for 3 hours.”

4.2 Larger Models

Our findings are based on using GPT-3.5-Turbo either directly (approaches 1,2,4,5) or on a fine-tuned variant of it (3). We believe these models have sufficient tacit knowledge in their weights to capture, in statistical terms, what Raskin (1984) and Raskin et al. (2009) refer to as *scripts*: a knowledge of the recurring events, rituals and practices that unite us as humans. For example, fine-tuning in approach (3) shows a recognition of a rather common annoyance in this novel output: “Why did the customer service representative cross the road? To talk to his manager.” LLMs unblock the representational bottleneck that has slowed progress in computational humour, and give joke generators wider scope over the gamut of human experience.

Larger LLMs, such as GPT-4, are capable of greater nuance in their analyses and of solving more complex problems (Achiam et al., 2023), so we might expect them to show more flair for joke generation too. When we retest 4 of our approaches (all but the fine-tuning of 3) on GPT4-Turbo, we obtain the results reported in Table 3.

Table 3: Breakdown of new and old jokes when GPT-4-Turbo is used over GPT-3.5-Turbo, with the corresponding rates of incoherence and sincerity in the outputs of different approaches

Approach	New Jokes	Old Jokes	Incoherent	Sincere
1. One-shot priming	14%	3.75%	4.4%	10.3%
2. Length sampling	22.59%	15.06%	4.6%	5.9%
4. RAG (two-shot)	15.11%	30.94%	3.6%	1.4%
5. CoT 6-step script	7.2%	0%	1.4%	0%

It appears, however, that GPT4 has an even greater propensity for retelling existing jokes. Approach 4, RAG (retrieval-augmented generation with two exemplar jokes) remains the most productive strategy, but it now leans more heavily on familiar favourites, or on near variants of the old reliables identified by Jentzsch & Kersting

(2023). As each joke exemplar is likely already known to the LLM, GPT-4 may take this as a license to lean into the familiar. However, its innate verbosity can also lead it to elaborate and stretch familiar jokes with short punchlines, as in this one-two punch: “How does a taco say grace? Lettuce pray [and may the forks be with you.]” The additional pun is not itself the punchline to another joke, but rather a quote from a *Star Wars* parody on *The Simpsons* TV show. The LLM has a magpie’s eye for shiny trinkets, and aptly matches the pseudo-religious *Force* pun with the wilted lettuce gag. Approach 2, sampling for concision and balance, appears to yield the best results for GPT-4, but its outputs remain very similar to those of GPT-3.5.

Chain-of-thought scripting (approach 5) remains the least effective strategy when using GPT4, as it tends toward verbosity, and so produces long setups with often meandering responses that attenuate the punch. Its outputs are affably charming in the jocular style of a late-night TV host, but without a topical context to spin, its outputs are best characterized as sardonic banter, not jokes of the *setup:punchline* variety. Still, these quips are very coherent, as reported in Table 3, and most show the requisite amount of pragmatic insincerity. This is insincerity that is meant to be penetrated by its intended audience (Kumon-Nakamura and Glucksberg, 1995) and approach 4 (RAG) also improves on this score (over GPT3.5-Turbo) with GPT4.

5. Conclusions

5.1 Mimics, Not Parrots; Magpies, Not Thieves

Skepticism toward new technologies is often warranted. Bender et al. (2021) urge us to be wary of the promise of large language models, noting that LLMs often behave like “stochastic parrots.” Here we have observed the tendency of specific approaches – particularly 2 and 3 – to parrot existing jokes from the LLM’s training data, and few who explore ChatGPT’s capacity for humour can fail to observe its tendency to produce endless jokes on the theme of chickens crossing roads (Jentzsch & Kersting, 2023). We must concede, however, that when an LLM is intelligently prompted it shows a strong feel for the inherent flexibility of familiar words to take on new and witty meanings.

Creative prompts can trick an LLM into coughing up large verbatim gobbets of its training data (Nasr et al., 2023), but we do not view regurgitation of this kind as a pressing concern. Our retrieval-augmented approach (4) is the most productive on GPT3-Turbo, and only 3.1% of its outputs are copies or close variants of existing jokes. So it is perhaps more accurate to call these generative systems “stochastic magpies.” These opportunistic magpies glean humorous conceits wherever they can, and repackage them as jokes with a classic setup and punchline shape, under the guidance of external pressures such as sampling for concision and balance. These conceits can come from anywhere, and may include typos or terms used in extended senses. The term “auto-pilates,” for instance, is used in several places on the web, either as a typo (of “autopilots”) or to denote the kind of Pilates one can do with a machine or in one’s car. It has been repurposed by approach 2 (sampling, GPT3.5) as follows: “How do self-driving cars stay in shape? By doing auto-pilates!” Silly

mistakes are grist for controlled silliness if an approach also encourages the use of what Giora et al. (2004) call *optimal innovations*. For example, the phrase “cover-up band” is mistakenly used on the web instead of “cover band,” but it acts as grist for approach 4 (RAG, GPT3.5) to generate this novel and memorable joke: “What do you call a group of conspiracy theorists? A cover-up band.” Optimal innovations that are coined by others, such as one website’s naming suggestions for hackers, can likewise be repurposed as jokes. Approach 4 reworks one such name for this joke: “What do you call a group of hackers? A CTRL-ALT-ELITE.”

5.2 To Find Sense in Nonsense, There is Method in Madness

There are reasons to be sceptical of, yet also optimistic about, LLMs developing a human sense of humour (Veale, 2021; Góes et al., 2023; Jentzsch & Kersting, 2023). GPT3.5-Turbo and GPT4-Turbo each bring a surfeit of training data, and no little raw talent, to the task of joke creation. Indeed, some of the jokes we observe from different approaches seem genuinely creative, and display a keen sense for the value of nonsense, as in this joke from the fine-tuned model of approach 3: “What did the keyboard say after a long day? What a bunch of kljdfhalkjds iuoldfhgaf.” At the same time, however, these approaches are also capable of generating truly execrable puns (such as “pun-dge” for a joking judge) that fall well below the standard set by older symbolic systems (Pain et al., 1997). Our findings suggest it not the size of a model that matters – although that can easily change – but how its outputs are coaxed and sampled by bespoke processes that model the comedic sensibility. For now, it is necessary to replicate the current findings on a swathe of different LLMs, both open and closed and of varying sizes, to identify some general insights into the capacity of these models to use words in ways that are deliberately humorous.

References

- Achiam, J., Adler, S., Agarwal, S., Ahmad, L. et al. (2023). GPT-4 Technical Report. *ArXiv* 2303.08774.
- Attardo, S., & Raskin, V. (1991). Script theory revis(it)ed: joke similarity and joke representational model. *Humor: International Journal of Humor Research*, 4(3), 293–347.
- Attardo, S. (1994). *Linguistic theories of humor*. Mouton de Gruyter, New York/Berlin.
- Attardo, S., Hempelmann, C. F., & Di Maio, S. (2002). Script oppositions and logical mechanisms: Modeling incongruities and their resolutions. *Humor: International Journal of Humor Research*, 15(1), 1–36. <https://doi.org/10.1515/humr.2002.004>
- Bender, E., Gebru, T., McMillan-Major, A., & Mitchell, M. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In A. Kasirzadeh, & A. Smart (Eds.), *Proceedings of FAccT-21 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–621).
- Coulson, S. (2001). *Semantic Leaps: Frame-Shifting and Conceptual Blending in Meaning Construction*. Cambridge University Press, Cambridge, UK.

Coulson, S., & Kutas, M. (2001). Getting it: human event-related brain response to jokes in good and poor comprehenders. *Neuroscience Letters*, 316(2), 71–74. [https://doi.org/10.1016/S0304-3940\(01\)02387-4](https://doi.org/10.1016/S0304-3940(01)02387-4)

Dean, G. (2001). *Step-by-Step to Stand-Up Comedy*. Heinemann, New Hampshire.

Evans, J. (2003). In two minds: dual-process accounts of reasoning. In two minds: dual-process accounts of reasoning. *Trends in Cognitive Sciences*, 7(10), 454–459. <https://doi.org/10.1016/j.tics.2003.08.01>

Freud, S. (1905). *Jokes and Their Relation to the Unconscious* (translated by J. Strachey). W. W. Norton, New York.

Giora, R., Fein, O., Ganzi, J., Levi, N. A., & Sabah, H. (2004). Weapons of Mass Distraction: Optimal Innovation and Pleasure Ratings. *Metaphor and Symbol*, 19(2), 115–141. https://doi.org/10.1207/s15327868ms1902_2

Góes, L., Sawicki, P., Grzes, M., Brown, D., & Volpe, M. (2023). Is GPT-4 Good Enough to Evaluate Jokes? In A. Pease, J. M. Cunha, M. Ackerman, & D. Brown (Eds.), *Proceedings of the 14th International Conference on Computational Creativity*, Waterloo, Canada.

Jentzsch, S., & Kersting, K. (2023). ChatGPT is fun, but it is not funny! Humor is still challenging Large Language Models. *ArXiv* 2306.04563. <https://doi.org/10.48550/arXiv.2306.04563>

Kahneman, D. (2011). *Thinking, fast and slow*. Penguin Books, London.

Kumon-Nakamura, S., & Glucksberg, S. (1995). How about another piece of pie: The allusional pretense theory of discourse irony. *Journal of Experimental Psychology: General*, 124(1), 3–21. <https://doi.org/10.1037/0096-3445.124.1.3>

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Sys*, 33, 9459–74. <https://doi.org/10.48550/arXiv.2005.11401>

Mihalcea, R., Strapparava, C., & Pulman, S. (2010). Computational Models for Incongruity Detection in Humour. In A. Gelbukh (Ed.), *Proceedings of the 11th CICLing conference on Computational Linguistics and Intelligent Text Processing* (pp. 364–374), Iasi, Romania.

Nasr, M., Carlini, N., Hayase, J., Jagielski, M., Cooper, A.F. Ippolito, D., Choquette-Choo, C.A. Wallace, E., Tramèr, F., & Lee, K. (2023). Scalable Extraction of Training Data from (Production) Language Models. *ArXiv* 2311.17035. <https://doi.org/10.48550/arXiv.2311.17035>

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L. et al. (2023). Training language models to follow instructions with human feedback. In Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Proceedings of the 36th Conference on Neural Information Processing Systems*, New Orleans, Louisiana.

Pain, H., Binsted, K., & Ritchie, G. (1997). Children's evaluation of computer-generated punning riddles. *Pragmatics and Cognition*, 5(2), 305–354. <https://doi.org/10.1075/pc.5.2.06bin>

- Pollio, H. R. (1996). *Metaphor: Implications and Applications*. In J. S. Mio, & A. N. Katz (Eds.), *Boundaries in Humor and Metaphor*, Lawrence Erlbaum Associates, New Jersey.
- Provine, R. R. (2000). *Laughter: A Scientific Investigation*. Viking Books, New York.
- Raskin, V. (1984). *Semantic Mechanisms of Humor*. D. Reidel, Dordrecht.
- Raskin, V., Hempelmann, C. F., & Rayz, J.M. (2009). How to understand and assess a theory: The evolution of the SSTH into the GTVH and the OSTH. *Journal of Literary Theory*, 3(2), 285–311. <https://doi.org/10.1515/JLT.2009.016>
- Reyes, A., Rosso, P., & Veale, T. (2013). A multidimensional approach for detecting irony in Twitter. *Language Resources and Evaluation*, 47(1), 1–30. <https://doi.org/10.1007/s10579-012-9196-x>
- Scott, S., Lavan, N., Chen, S., & McGettigan, C. (2014). The social life of laughter. *Trends in Cognitive Science*, 18(12), 618–620. <https://doi.org/10.1016/j.tics.2014.09.002>
- Shahaf, D., Horvitz, E., & Mankoff, R. (2015). Inside Jokes: Identifying Humorous Cartoon Captions. In T. Joachims, G. Webb, D. Margineantu, & G. Williams (Eds.), *Proceedings of the 21st ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (pp. 1065–74), Sydney, Australia.
- Toplyn, J. (2014). *Comedy Writing for Late-Night TV*. Twenty Lane Media, New York.
- Toplyn, J. (2022). Witscript-2: A System for Generating Improvised Jokes Without Wordplay. In M. Hedblom, A. Kantosalo, R. Confalonieri, O. Kutz, & T. Veale (Eds.), *Proceedings of the 13th International Conference on Computational Creativity*, Bozen, Italy.
- Veale, T. (2021). *Your Wit Is My Command: Building AIs with a sense of humor*. MIT Press, Cambridge MA.
- Wei, J., Xuezhong W., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Proceedings of the 36th Conference on Neural Information Processing Systems*, New Orleans, Louisiana.
- West, R., & Horvitz, E. (2019). Reverse-engineering satire, or “paper on computational humor accepted despite making serious advances.” In B. Williams, Y. Chen, & J. Neville (Eds.), *Proceedings of the AAAI Conference on Artificial Intelligence* (pp. 7265–72), Honolulu, Hawaii.
- Winters, T. (2023). *Neuro-Symbolic Creative Artificial Intelligence for Humor*. [Doctoral dissertation, University of K.U. Leuven]. <https://lirias.kuleuven.be/retrieve/736717>

Contact information

Tony Veale

School of Computer Science, University College Dublin, Ireland
tony.veale@UCD.ie

Valeria Caruso and Lucia di Pace

WORDS FOR CHOOSING FOOD IN ALMA – MULTIMEDIA ATLAS OF BIO/CULTURAL FOOD

Abstract This paper presents an innovative lexicographic approach embedded within an online resource currently under development: *ALMA: Linguistic Multimedia Atlas of Bio/Cultural Food Diversity*. *ALMA* serves a dual purpose: firstly, to showcase linguistic diversity through culinary practices, and secondly to scrutinise food marketing strategies through the analysis of language and paralinguistics on food packaging. The resource has two main components: *Words for Food* (WF) and *Words for Choosing Food* (WCF) lexicographic articles. The WF articles delve into the lexicon of common language and technicalities of food practices, recognizing the dynamic interaction between common and specialised knowledge of food production. On the other hand, the WCF articles employ a segmentation methodology to scrutinize food packaging, identifying functional and linguistic constituents.

By summarizing individual labels and pinpointing issues regarding lack of transparency, the WCF articles facilitate consumer understanding and stimulate debates on labelling standards. This innovative lexicographic approach aims to empower consumers with linguistic and encyclopaedic knowledge while providing insights for legislators and manufacturers to enhance food labelling practice.

Keywords specialised language; food language; online lexicographic resources; food labels corpora; legal terminology; consumer protection texts

1. Introduction

This study presents the lexicographic articles of *ALMA: Linguistic Multimedia Atlas of Bio/Cultural Food Diversity* (Caruso & di Pace, 2024). The name *ALMA* has multiple meanings and evokes many components of the resource, as it stands for ‘nutrient’ in Latin, ‘soul’ in Old Italian, and ‘apple’ in Hungarian.

Currently in development, this online tool serves a dual documentary purpose. Primarily, it seeks to spread language knowledge through culinary practices by showcasing the local cultural heritage of renowned areas, celebrated for their traditions. At the same time, it explores various facets of food marketing, examining the language and paralinguistics used on packaging.

The linguistic choices crafted for sales purposes are explained in dedicated terminology articles which analyse all the components of the product label:

- the different product names, i.e., *legal*, *customary* or *descriptive* (EU Regulation, n. 1169/2011)¹;

¹ The regulation is retained also in the UK under the EU (Withdrawal) Act 2018.

- all mandatory and non-mandatory textual elements, e.g., list of ingredients, storage conditions and/or conditions of use;
- graphic components, e.g., logos and pictograms.

This part of the resource aims to analyse the *Words for Choosing Food* (WCF), complementing the description of *Words for Food* (WF). The latter pertains not only to the lexicon of common language but also to the technicalities of food practices, acknowledging the fluid contexts of food production.

Following a concise presentation of the resource, the paper introduces the innovative aspects of the WCF section (Figure 1), which is based on a corpus of food labels created with the segmentation methodology used by Wiegand for dictionaries (Wiegand, 1984; Wiegand & Smit, 2013, § 3). According to this procedure, food packaging is segmented into macro-components, such as pack faces or different labels on the packaging. Within these components, various functional constituents and their specific *items* are identified, revealing several elements of interest in the terminology employed by the producers.

The two components of the resource aim to provide users with a variety of information on food, enhancing their linguistic and encyclopaedic knowledge, as well as their understanding of food labels. Similar to other consumer protection texts (e.g., pharmaceutical leaflets, see di Pace, 2019), food labels require specific skills to interpret. In this context, the WCF lexicographic entries summarize the contents of individual labels, highlighting elements of lack of transparency that may obscure product features.

This approach should also stimulate reflection on labelling standards and practices, providing insights for both legislators and manufacturers. Currently, mobile applications, such as Yuka (Yuka App), are available to provide guidance on the healthiness of packaged foods and assist consumers in making purchase decisions based on health-related motivations. However, consumers also choose food products based on other parameters, such as taste, ethics, or environmental friendliness and manufacturers use various types of information to influence consumers' choices, some of which may be hidden.

For all the purposes described above, a lexicographic approach appears effective in providing information quickly and efficiently through dictionary articles for any interested user. Furthermore, the use of general language dictionaries is recommended in the food labelling process, as they serve as a strategic reference by offering a reliable repository of encyclopaedic information shared by a language community.

2. ALMA Components

Recent research has mostly focused on analysing the language of food from an interactional perspective, with studies such as Rüdiger & Mühleisen (2020) and Tovares & Gordon (2020) demonstrating how social status is negotiated through food.

In contrast, studies conducted in previous years by Silvestri (see *inter alia* Silvestri, 1999; 2002) have outlined a comprehensive method of analysis to explore the language of food from micro- to macro-linguistic levels. These studies represent the foundation upon which *ALMA* is currently developing, by also implementing the descriptions proposed by Silvestri with multimedia content to illustrate various aspects of food-related multimodality lexicalised by the WF, such as food shapes, colours, processing procedures and rituals connected to the food sphere.

The descriptive potential of *ALMA* and the richness of the data it provides for further research emerges by illustrating the lexicographic articles of the resource. Lemmas like “pan bauletto” (Figure 1), for instance, allow the user to access one article of the *Words for Food* section, which describes linguistic properties of words and technical terms. At the same time the user is referred to various articles of the *Words for Choosing Food* section, which are attached to the same type of aliment, e.g., the Italian *panbauletto* in Figure 1, despite the wide variety of selling names used by the manufacturers. More details about the lemmatisation in use for the WCF articles will be provided in dedicated paragraphs (§ 3.2).

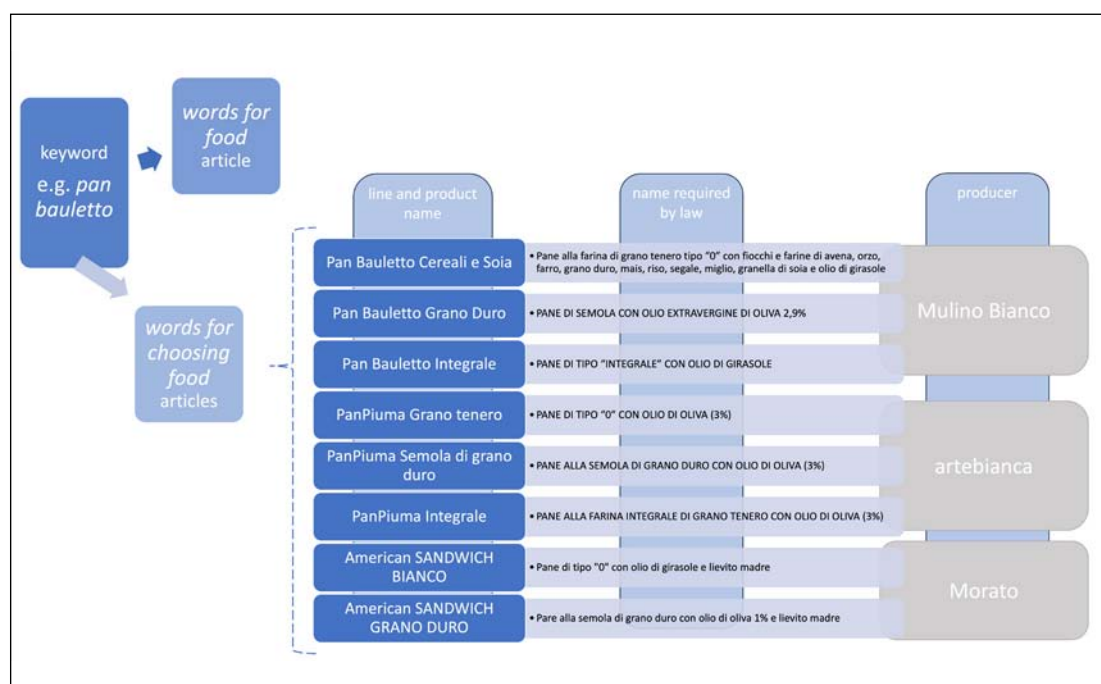


Fig. 1: General structure of *ALMA*. Food terms are described by two types of lexicographic articles: one, called *Words for Food*, provides lexical descriptions, and the other, called *Words for Choosing Food*, focuses on food labels. Both can be queried based on the type of product, regardless of their commercial names or the names required by law used for food products on the labels.

2.1 Words for Choosing Food

In the *ALMA* section dedicated to the *Words for Choosing Food*, linguistic and paralinguistic aspects emerging from the analysis of packaged food labels are collected. Within the EU, food products must comply with all obligations imposed by a specific legislation which defines ‘prepacked food’ as:

any single item for presentation as such to the final consumer and to mass caterers, consisting of a food and the packaging into which it was put before being offered for sale, whether such packaging encloses the food completely or only partially, but in any event in such a way that the contents cannot be altered without opening or changing the packaging (EU 1169/2011, art. 2, 2i).

It should be noted that the literature on labels has developed extensively in recent years, both in dedicated monographs (for example, see Albert, 2014; Berryman, 2015; Smith et al., 2022) and in scientific articles published in numerous journals such as the *Journal of Consumer Policy*, the *Journal of Consumer Protection and Food Safety*, *Appetite*, and *Food Quality and Preference*. The literature primarily focuses on issues of health protection or market constraints, covering fields of study that are social or, more specifically, micro-economic. However, studies dedicated to the linguistic analysis of labels seem to be lacking, despite the fact that labels are multimodal macro-texts consisting of names, texts, symbols, images, and more.

This deficiency is particularly notable considering that even the simple naming of foods is subject to debates and conflicts, driven by the need to protect traditional products of high quality and prestige. For instance, consider the legal process for registering *Pizza Napoletana* as a Traditional Speciality Guaranteed (TSG) with name reservation (EU Regulation 2022/2313). However, the issue of how to name foods remains one of the most confusing aspects within the legislative framework.

Therefore, the objective of the *ALMA* project is to thoroughly analyse the various components of labelling to highlight a) the different product denominations and their types; b) the use of marketing strategies to influence purchasing decisions; c) the lack of transparency inherent in the information provided to the consumers.

2.2 Words for Food

The features of *ALMA* are illustrated here by means of concrete examples, beginning with a case study: the lexicographic description of a popular Italian prepackaged bread, *Pan Bauletto Bianco*, produced by a major manufacturer in the country. The commercial name *Pan Bauletto* also gave rise to a neologism, *panbauletto*, since the product became very popular as the basis for many recipes, although it has not yet been lemmatized in dictionaries. The *Pan Bauletto Words for Food* article (Figure 2) allows for a better understanding of the product's characteristics and provides a quick overview of the *ALMA* articles on common and technical terms used in the food sector. This bread, which resembles the traditional English loaf, takes its name from the curved shape of the top of the slices, which makes it resemble a portmanteau, or *baule* in Italian.

The WF articles collect the data necessary to understand the named objects and also complement the information provided in the WCF articles on food labelling. This information is particularly relevant for clarifying the nature of products containing ingredients used in industry but not common in homemade preparations, such as *malt flour*, *gluten*, or *inverted sugar*, to name a few. Many of these ingredients, as well

as food additives, are rarely known to consumers, and *ALMA* aims to offer prompt support regarding them to its users.

	Lemma	Pan Bauletto * words for choosing food variants: <i>panbauletto</i> (graphic variant due to univerbation), <i>pane bauletto</i> (lexical variant, without loss of the final vowel -e)
	Type	Products; Bread. Ordinary language
	Morphology	<i>Pan Bauletto</i> , compound word <i>Pan</i> (shortening of <i>pane</i> n. m. 'bread') <i>Bauletto</i> (n. m., derived word on the base <i>baule</i> 'case' + evaluative suffix -etto, with the meaning 'little case')
	Semantics	Metaphor: the shape of the loaf from the mould in which bread is baked A very soft type of loaf, made with olive oil, with an elongated, rounded shape resembling that of a trunk. It is widely used for stuffed sandwiches, savoury snacks and as a basic ingredient in many recipes.
	Meaning	<i>Io, come te, faccio il pane a lievitazione naturale, a volte tipo pan bauletto</i> <i>Ho fatto anche un pan bauletto più dolce</i> <i>dividere a metà la salsiccia, schiacciare per bene sulla fetta di pan bauletto</i> <i>Sopra il panbauletto mettete la crema di ricotta</i> <i>Mi piace molto questo pane bauletto di farro integrale biologico</i> <i>Il pane bauletto è un ottimo sostituto del pane fresco</i>
	Hypernym	<i>pane in cassetta</i> , multi-word, 'bread baked in a box', 'sandwich bread', 'pullman bread'
	Co-hyponym	<i>Pancarré</i> , compound n. m. : <i>pan(e)</i> + <i>carré</i> adj. 'squared' (false French loanword)
	Source	Written sources: corpora; website of bread company; advertising images; bread packs
	Previous and following levels	-1 dough; 2 bruschetta
	Etymology	From a Product Name, originally registered by ESKA Snc Di Scucchiaro Lino & Co (19/10/2007 UIBM), now produced by Mulino Bianco (trademark)
	Ethno-socio-psycholinguistic implications	"Attenzione, il pancarré è superato; adesso si tosta una cosa chiamata Pan Bauletto"
	Notes	

Fig. 2: *Pan Bauletto* Word for Food article in *ALMA* translated into English from the original Italian article

To effectively capture the characteristics of food products, an ontological model developed specifically for these types of concepts by Silvestri in various studies is used (see *inter alia* Silvestri, 1999; 2002). Foods are considered as “products” within a supply chain defined by different “procedures” representing the work required to produce and consume them. This methodology traces production chains through different “processing stages,” from the production of ingredients to food consumption practices, including events such as meals; a sketch of this ontology is provided in Caruso & di Pace (2024).

Four main types of information are provided in the lexicographic entries of the *Words for Food* articles: i) formal aspects of the lemma (its variants, morphological description); ii) its semantics; iii) relevant phraseology; and iv) ethno-socio-anthropological implications.

The semantic description offered in the articles refers to the ontological organisation of *ALMA*’s data. It briefly presents the different ‘levels of processing’ (indicated as ‘previous and following levels’ in the form), the paradigmatic dimension (synonyms/antonyms, holonyms/meronyms), and the type of word, i.e., whether it lexicalizes a product or a process and whether the lexicalization is a metaphor, metonymy, or synecdoche. These specifications enable users to search the database to obtain lists of lemmas corresponding to the mentioned typologies.

3. A Lexicographic Method for Label Analysis

The three objectives outlined above – collecting different types of product names, understanding marketing strategies that influence consumer motivation in food purchases, and addressing the lack of transparency in the information provided to consumers – are achieved through a systematic study of food labels. This study is conducted using a lexicographic approach, which serves as the basis for creating WCF articles. These synthesise the characteristics of food labels on the basis of both the “mandatory” information required by law and the “voluntary” information provided by the manufacturer.

The method of analysis and description used is based on the structuralist decomposition method used for dictionaries by H. E. Wiegand (Wiegand, 1984; Wiegand & Smit, 2013). The first step in this procedure is to identify the macro-components to be analysed, in our case the “faces” of the packaging, which are then broken down into individual items, i.e., units with a specific function that can also be further segmented. Figure 3 illustrates this method with the segmentation of a *Pan Bauletto* pack sold in Italy, the back of which has been broken down into 16 items², some of which have been further subdivided into their constituent elements.

The item numbered as 3, for example, includes the product name required by law (indicated as 3a, see § 3.2) and the list of ingredients (3b), whose progressive number is registered and used as a reference to illustrate the characteristics of the packaging (see Table 1).

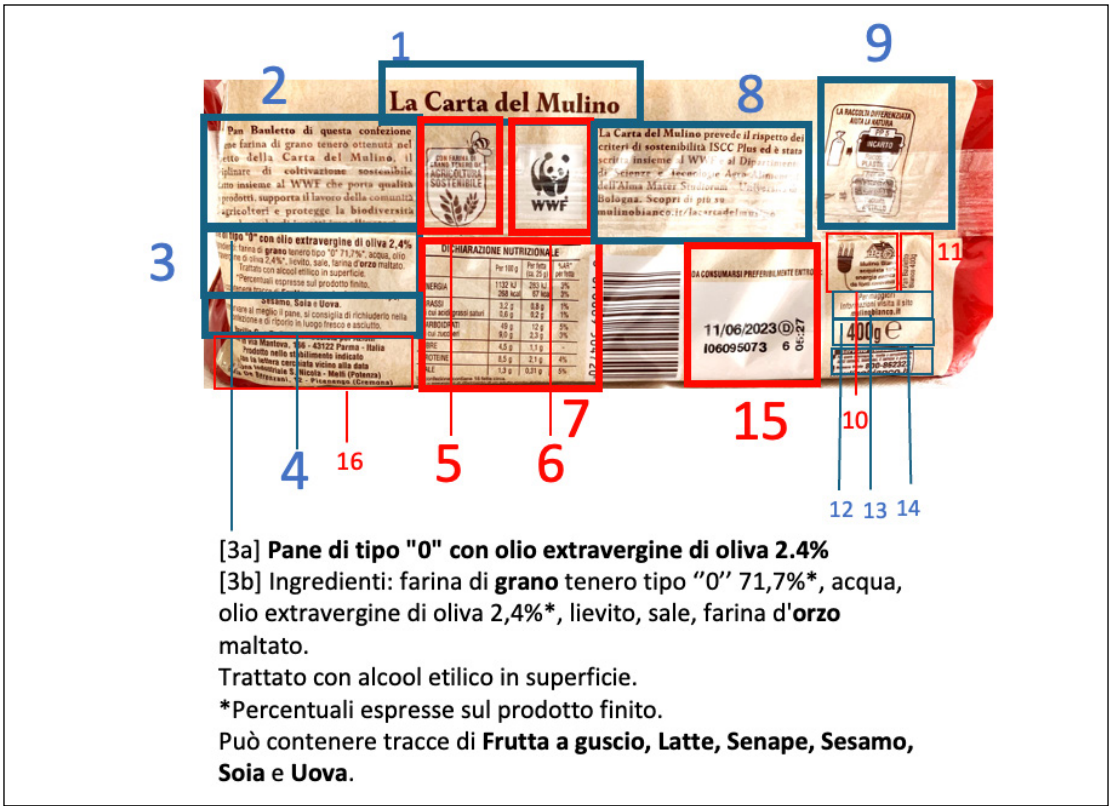


Fig. 3: Example of decomposition of a food label

² The terminology used here follows that introduced by Wiegand in his *General Theory of Lexicography* (see *inter alia* Wiegand, 1984; Wiegand & Smit, 2013).

Table 1: Segmented packaging *items* with the corresponding “purchase motivation” and aspects of lack of transparency (or “issues” in the form)

n°	element	Type	purchase motivation	side	issues	specification	notes
8	La Carta del Mulino prevede il rispetto prevede il rispetto dei criteri di sostenibilità ISCC Plus...	Text	natural content	back	obscurity		acronym “ISCC Plus”
3b; 18	Farina di grano tenero tipo “0”	ingredients	natural content	back	obscurity	bold type	do not explain that bold types highlight an allergen
3b; 27	Frutta a guscio	ingredients		back	obscurity	capital letter	anomalous use of capital letters
5	pictogram	logo with text	natural content	front	replication		
5	pictogram	logo with text	ethical content	front	replication		
2	Mulino Bianco	company name	ethical content	side2	replication	bee icon	
2	Ogni Pan Bauletto di questa confezione...	text		back	repetition	textual	2, 8
2	Ogni Pan Bauletto di questa confezione...	text	ethical content	back	repetition	inter-semiotic	2, 6, 8
3b; 21	lievito	ingredients		back	vagueness		which yeast type?

The *item* collection provides a basis for outlining the label features according to the dimensions listed above (points: § 2.1, a, b, c). These features are supplemented in a lexicographic article of the type exemplified in Figure 4.

The article offers linguistic data as the first piece of information, lemmatising the primary name of the product, which is generally a product name (or commercial name, Vaxelaire, 2014) used for marketing purposes displayed on the principal field of vision (or ‘front of pack’). A morphological segmentation follows, analysing the components of the name used in the industry (Forte & Della Penna, 2017, § 3.2). In the case of *Pan Bauletto Bianco*, the line name is *Pan Bauletto*, as many other types of this bread are produced, and the product is called simply *Bianco* (‘white’), which is a collocate standing for ‘wheat flour’. For a strictly morphological analysis based on linguistic principles, the user is referred to the general WF article on *Pan Bauletto*.

	Name on the Front of the Pack/Primary Name FOP	Pan Bauletto Bianco Line and Product name
	Morphology	<i>Pan Bauletto Bianco</i> , multi-word <i>pan bauletto</i> + words for food, n. m. (compound), line name <i>bianco</i> , adj., product name
	Semantics	<i>Pan Bauletto</i> ; <i>bauletto</i> = metaphor (from the shape) <i>Bianco</i> = with wheat flour
	Secondary Name BOP	Pane di tipo «0» con olio extravergine di oliva 2.4% , multi-word (legal name+specification) <i>Pane di tipo «0»</i> , legal name <i>con olio extravergine di oliva 2.4%</i> , specification of the additional ingredient
	Pack Items 48	Ingredients 16, Logo with text 4, Slogan 3, Line and product name 3, Text 3, Image 3, Company name 3, Pictogram with text 2, Product description 1, Manufacturer 1, Date 1, Title of text 1, References for the consumer 1, Name required by law 1, Nutrition table 1, Weight 1, Logo 1, Product identification 1, Storage conditions 1
Purchase motivation average per item 1,35		48 items, which mostly offer information related to health protection, the naturalness of the product and respect for ethical choices (through reference to specific agricultural disciplinary rules, drawn up in collaboration with public institutions and non-governmental organisations). none 34% health 19% ethical 18% natural content 18% sensory appeal 9% weight control 1%
Item Issues average per item 1,18		The label is characterised by redundancy (44% of the items, replication + repetition) and contains elements that may be difficult to understand (25%, obscurity + vagueness) none 43% replication 32% obscurity 23% repetition 12% vagueness 2%
Manufacturer		M*** B***

Fig. 4: An article from *the Word for Choosing Food* in *ALMA*: Pan Bauletto Bianco. The article has been translated into English from the original Italian.

Information on semantics provides a more comprehensive explanation of the product name and affords the lexicographer the opportunity to analyse marketing strategies in product naming. This aspect is not as thoroughly explored in the case of *Pan Bauletto Bianco*, as its naming relies on a simple metaphor, namely the word *bauletto* which is derived from the curved upper side of the bread. However, readers of this paper are well aware of many product names that stand out for the savvy associations they evoke. *Q-tips*, a famous product name for cotton swabs, is considered to have gained commercial success after its naming, with the “Q” standing for ‘quality’ and perhaps subliminally evoking the shape of ears.

The secondary name, typically printed before the ingredients list, is the name required by law, as it will be explained in a moment (§ 3.2). The other sections of this article point towards describing the richness of the packaging contents, providing a count of the *items* as indicated by the “Pack items” search zone³ (48 on the label of *Pan Bauletto Bianco*, see Figure 4). They also highlight the producers’ efforts to pursue special marketing strategies by indicating “Purchase motivation[s]” (see the corresponding search zone in the article), which can be fostered by specific label features, and the issues (“Item issues” search zone) that users can experience when interacting with the label. These issues may hinder consumers’ ability to make informed purchase choices, especially with regard to elements of obscurity, vagueness, and ambiguity.

³ Search zones in dictionary articles are aggregated data blocks facilitating the look up process, see Gouws (2014).

3.1 Motivation in the Choice of Foodstuff

The different purposes for which producers decide to present certain contents on labels align with the motivations behind dietary choices of consumer segments (“Purchase motivation” search zone). These motivations are discussed extensively in the literature (starting with Steptoe, Pollard, & Wardle, 1995), which highlights how consumers may be guided by considerations related to: a) health (seeking proper nutrition or foods considered healthier than others); b) *natural content* (preference for foods free of pesticides, preservatives, etc.); c) sensory appeal (choice of food not only on the basis of taste but also on the basis of smell and appearance); d) weight control (choice on the basis of calories information, etc.); e) familiarity (tendency to prefer what is perceived as familiar rather than new); f) ethic concern (environmental concerns, e.g., choice of not eating meat, but also preference for food packaged in recyclable materials); g) convenience; h) price; i) mood (as systematised by Cavazza & Guidetti, 2020, pp. 15–16).

In analysing the *Words For Choosing Food*, each *item* on the label is linked to a specific purchase motivation. Hence, the term “Words For Choosing Food” is employed, as all elements on food labels aim to guide consumer product choices.

3.2 Legal Names, Commercial Names and Lemmatisation Issues

The initial challenge in compiling concise descriptions of food labels in the form of lexicographic entries, as exemplified in Figure 4, arises from lemmatisation.

As illustrated in Figure 1, the product names of industrial food items are primarily determined by manufacturer-established principles (Forte & Della Penna, 2017). A specific name identifies the product line (e.g., *Pan Bauletto*), followed by specifications that can be used to distinguish different types of the same line: e.g., *Pan Bauletto Integrale*, *Pan Bauletto di Grano Duro*, *Pan Bauletto Cereali e Soia*.

However, in addition to these commercial names, there must be a name with legal value, which is the only one required to trade a food product in the European Community and the United Kingdom. There are, however, a number of critical aspects to be considered in relation to this legal designation.

First of all, there are three types of names by which a food can be designated. Their differences are due to: the existence of specific laws for certain types of products, the availability of common names in linguistic usage and, in the absence of the former, the need to designate foods clearly enough to enable consumers to “understand their nature” (EU 1169/2011, Art. 1, 2, p). The choice of one type of name over another is thus limited: only in the absence of the first type, or *legal name*, can one resort to the second, *customary name*, and thus to the third, *descriptive name*. The contradictory nature of the terminology introduced by the legislator must be emphasised, as all the proposed names hold legal value, yet the adjective *legal* is used to refer to only one of them (*the legal name*). In the absence of an appropriate hypernym to denote all three

types of names, we have opted for *name required by law* as the more general term in the context of this research.

3.3 Current Strategies in Pre-Packaged Food Naming

A preliminary analysis of a number of Italian products (Caruso & di Pace, in press) has made it possible to outline the practices used in the naming of foodstuffs in the country. Firstly, there is a multiplicity of names on the same packaging (sometimes up to five), secondly, there is a great variation in the *names required by law* of products from the same line and, finally, there are cases of poor fidelity to *realia* due to the misuse of common names.

This phenomenon is observed in the case of the ‘plum cakes’ of a well-known Italian company. This name is currently used in Italy for a very popular snack. In everyday usage, the term *plum cake* refers to a fluffy sponge treat made with yogurt rather than the traditional British ‘plum cake’. However, the *plum cakes* in this line are called *brioche* by the manufacturer in their *descriptive name* (or the *name required by law* in Table 2):

Table 2: Examples of the different designations used as *names required by law* of products from the same line, by the same manufacturer, compared to the *commercial name* and the additional designations used on the same pack

Commercial Name	Name Required by Law	Additional name
PlumCake Classico	Brioche preparata con yogurt	Merende
PlumCake con Granella di Mandorle e Ricotta	Brioche con mandorle e ricotta, senza zuccheri aggiunti, con edulcoranti con aggiunta di calcio	
PlumCake Integrale	Brioche preparata con farina integrale di frumento e yogurt	Merende

However, it is well known that brioche belong to the family of ‘enriched doughs’, as they are described in gastronomic glossaries or:

[a] dough made with butter, sugar, eggs or oil. Enriched doughs are softer and richer than those made solely with flour, water, salt and yeast. Examples include brioche and hot cross buns. (Baking Glossary)

Other questionable practices concern the use of modifiers within the descriptive name. Sometimes evaluative adjectives are used that one would not expect to find in a *name required by law*. For example, another Italian manufacturer’s ‘plum cake’ is a “Prodotto dolciario da forno. Merende di soffice Pasta Margherita preparata con Yogurt Magro” (En. ‘Bakery confectionery. Snacks made of soft sponge cake dough prepared with low-fat yoghurt’). Note in this detailed description the specification of the ‘fluffiness’ of the sponge cake and the different ways of referring to the snack cake, both as a “baked confectionery product” and the classifier “snacks made of.”

Likewise, the descriptive names coined by manufacturers are very complex text strings, sometimes small captions, which need to be properly described and segmented.

The anomalies of the names already illustrated, such as a yoghurt snack being described as an ‘enriched dough’, will be clarified for the end user of *ALMA* thanks to its mediostructure, which will allow hyperlinks between WCF and WF articles. The latter in fact contain the encyclopaedic information necessary for the user to fully understand any descriptive inaccuracy.

3.3.1 Redundancy

For a proper assessment of packaging features, the proliferation of synonyms referring to the same product is recorded as a case of “redundancy.” Other elements contributing to this effect are “repetitions” and “duplicates”, which can also occur in an intersemiotic dimension (see Jakobson, 1959) when logos, images, or pictograms are used that refer to the same contents presented in discursive form within texts.

On the packaging synthesised by the entry in Figure 1, two texts (item 2 and item 8 in Figure 3) appear emphasising how the company adheres to a discipline that promotes sustainable agriculture and ethical principles:

Each Pan Bauletto in this pack contains wheat flour obtained in accordance with the [Manufacturer’s name] Charter, the sustainable cultivation specification drawn up together with the WWF, which brings quality to the products, supports the work of the farming community, and protects biodiversity by also encouraging pollinating insects. (text of *item 2* on the *Pan Bauletto* pack in Fig. 3)

Logos 5 and 6 communicate the same principles, and the image of the wheat ear on the front of the pack (reproduced in Figure 4) reaffirms the same ethical content. Furthermore, logo 5 is replicated on three different faces of the pack.

These aspects of redundancy represent one of the descriptive parameters used in the WCF entry, listed alongside aspects of lack of transparency among the “Item issues.” A mean is provided for these elements: in the case of *Pan Bauletto Bianco*, an average of 1.18 “Item issues” and of 1.35 “Purchase motivation[s]” per *item* has been identified.

3.3.2 Lack of Transparency

The elements that lack transparency are also registered, illustrated and classified as cases of “vagueness”, “ambiguity” or “obscurity.” In this context, we define “transparency” as the ability of a text to be accurately interpreted by its intended user. Vagueness occurs in all cases where the meaning of the term is not defined and circumscribed (Cruse, 1986, p. 81; Ježek, 2005, p. 70). For example, classificatory adjectives such as *high/low*; *young/old*; *beautiful/ugly* are generally considered to be vague because there are no explicit parameters or scales of reference to guide their use. Similar examples are common on food packaging, especially regarding storage instructions. For packaged bread, it is generally stated: “Store in a cold, dry place, away from sunlight”, but it is not clear how many degrees guarantee that a place is sufficiently “cold” or what conditions determine its dryness.

In contrast, a term is “ambiguous” when several meanings are legitimate (see also Cruse, 1986, pp. 49–68). For example, to the Italian noun phrase *zucchero cristallino*⁴ two different values are ascribed in general language dictionaries. For De Mauro, the multi-word has the technical-scientific meaning of “solid matter that appears in the form of a crystal or an aggregate of crystals: *crystalline sugar*, *crystalline ice*.” The Treccani, on the other hand, claims a ‘commercial’ use to designate the product of the “first crystallisation [of sugar], purified and dried, which is produced for industrial use”, and which is distinct from “refined or granulated sugar.”

Obscurity, on the other hand, concerns expressions that cannot be interpreted, either because of information gaps in the text itself or because of insufficient knowledge on the part of the interpreter/consumer. In this regard, for *ALMA*, we opted to utilize general language dictionaries as a reference point to evaluate the accessibility of information for consumers. In instances involving terminological neologisms, corpora were consulted for verification. For example, the term *panbauletto* discussed earlier is currently in use with the meaning described in the WF article of Figure 4, as evidenced by corpus attestations (cfr. Caruso & di Pace, 2024).

4. Concluding Remarks

This paper has outlined the features of a new type of lexicographic article intended to provide a concise yet explanatory description of food labels.

The proposal aims to address consumers’ information needs in alignment with the purposes indicated by the most recent legislation in effect in the EU.

It is expected that the analysis, based on lexicographic principles, will enhance consumers’ understanding of food labelling and assist legislators in refining regulations. Consumers are presented with concise descriptions – such as those in Table 1 – of label components aligning with different purchase motivations and others that lack transparency. Furthermore, they are presented with specific linguistic analysis and encyclopaedic information to acquire better knowledge on food topics and make better informed purchase choices.

Lastly, legislative terminology and the analysis of food designations appearing on labels can be the subject of more specialised analysis in dedicated publications (Caruso & di Pace, in press).

References

Albert, J. (Ed.). (2014). *Innovations in Food Labelling*. Elsevier – Woodhead Publishing.

Baking Glossary, *Baking As Lifestyle Medicine Glossary*. Retrieved June 2, 2024, from <https://www.sourdough.co.uk/glossary/enriched-dough/#:~:text=A%20dough%20made%20with%20butter,brioche%20and%20hot%20cross%20buns>

⁴ The noun phrase appears in a descriptive name such as “Croissant con zucchero cristallino in superficie”, En. ‘croissant with crystalline sugar on the surface’.

Berryman, P. (Ed.). (2015). *Advances in Food and Beverage Labelling. Information and Regulations*. Elsevier – Woodhead Publishing.

Caruso, V., & di Pace, L. (2024). I saperi alimentari e la loro descrizione linguistica e multimodale. Esempi da *ALMA: Atlante Linguistico Multimediale della diversità Alimentare bio/culturale*. In L. Cirillo, & R. Nodari (Eds.), *Contesti, pratiche e risorse della comunicazione multimodale* (pp. 45–63). Studi AITLA.

Caruso, V., & di Pace, L. (in press). Analisi linguistica delle etichette alimentari: la questione delle denominazioni tra circolarità normativa e vaghezza terminologica. In J. Altmanova, M. Centrella, & C. Grimaldi (Eds.), *Terminologia del patrimonio culturale materiale e immateriale: analisi e approcci di studio*. Peter Lang.

Cavazza, N., & Guidetti, M. (2020). *Scelte alimentari*. Il Mulino.

Cruse, D. A. (1986). *Lexical Semantics*. Cambridge Textbooks in Linguistics.

De Mauro, *Il nuovo De Mauro*. Retrieved June 2, 2024, from <https://dizionario.internazionale.it/>

di Pace, L. (2019). *La lingua del bugiardino. Il foglietto illustrativo tra linguaggio specialistico e linguaggio comune*. Franco Cesati Editore.

Forte, M., & Della Penna, A. (2017). *L'Oréal e la Vache qui Rit*. Analisi semio-linguistica di due nomi commerciali Francesi. *Lingue e Linguaggi*, 22, 81–96. DOI 10.1285/i22390359v22p81

Gouws, R. H. (2014). Article Structures: Moving from Printed to e-Dictionaries. *Lexikos*, 24, 155–177. <https://doi.org/10.5788/24-1-1256>

Jakobson, R. (1959). On Linguistic Aspects of Translation. In R. Brower (Ed.), *On Translation* (pp. 232–239). Harvard University Press.

Ježek, E. (2005). *Lessico. Classi di parole, strutture, combinazioni*. Il Mulino.

REGULATION (EU) No 1169/2011 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 25 October 2011 on the provision of food information to consumers.

REGULATION (EU) No 2022/2313 OF COMMISSION IMPLEMENTING REGULATION (EU) of 25 November 2022 entering a name in the register of traditional specialities guaranteed (Pizza Napoletana (TSG).

Rüdiger, S., & Mühleisen, S. (Eds.). (2020). *Talking about Food: The social and the global in eating communities*. John Benjamins Publishing Company.

Silvestri, D. (1999). Demografia, sistemi agrari, regimi alimentari nel mondo antico. In D. Vera (Ed.), *Atti del Convegno Internazionale di Studi (Parma 17-19 ottobre 1997)* (pp. 345–365). Edipuglia.

Silvestri, D. (2002). Introduzione. In D. Silvestri (Ed.), *Apriti, sesamo! “Bollettino dell’Atlante Generale dell’Alimentazione Mediterranea* (pp. 9–13). Quaderni di AION n.s. 4. Il Torcoliere.

Smith, V., Barratt, D., Møgelvang-Hansen, P., & Wedel Andersen, A. U. (2022). *Misleading Marketing Communication. Assessing the Impact of Potentially Deceptive Food Labelling on Consumer Behaviour*. Palgrave MacMillan.

Steptoe, A., Pollard, T. M., & Wardle, J. (1995). Development of a Measure of the Motives Underlying the Selection of Food: The Food Choice Questionnaire. *Appetite*, 25(3), 265–284. <https://doi.org/10.1006/appe.1995.0061>

Tovares, A., & Gordon, C. (Eds.). (2020). *Identity and ideology in digital food discourse: Social media interactions across cultural contexts*. Bloomsbury Academic.

Vaxelaire J.-L. (2014). Point de vue sur les noms déposés. In M. Van Campenhoudt, N. Lemaire, & R. Temmerman (Eds.), *Traduire aux confins du lexique: les nouveaux terrains de la terminologie. Cahiers de linguistique. Revue de sociolinguistique et de sociologie de la langue française* 40(2), 159–176.

Vocabolario Treccani online. Retrieved June 2, 2024, from <https://www.treccani.it/vocabolario/>

Wiegand, H. E. (1984). On the Structure and Contents of a general Theory of Lexicography. In R. R. K. Hartmann et al. (Eds.), *LEXeter '83 Proceedings* (pp. 13–30). Max Niemeyer.

Wiegand, H. E., & Smit, M. (2013). Microstructures in printed dictionaries. In R. H. Gouws, U. Heid, W. Schweickard, & H. E. Wiegand (Eds.), *Dictionaries. An International Encyclopedia of Lexicography: Supplementary volume. Recent Developments with Special Focus on Computational Lexicography* (pp. 149–214). Mouton de Gruyter.

Yuka App, *Yuka – Food & Cosmetic Scanner*, Yuca, SAS, Paris: France <https://yuka.io/>, last update May 2024.

Acknowledgements

This work is part of a research project funded by the Department of Literary, Linguistic, and Comparative Studies at the University of Naples “L’Orientale”: *ALMA: Multimedia Linguistic Atlas of Bio/Cultural Food Diversity*, coordinated by Anna De Meo (DR No. 0007984 of 02/01/2022, Rep. 75/2022).

This work is the result of continuous collaboration between the authors. However, Valeria Caruso is responsible for paragraphs 2, 2.2, 3, 3.3.1, 3.3.2, while Lucia di Pace is responsible for paragraphs 1, 3.1, 3.2, 3.3, 4.

Contact information

Valeria Caruso

Università di Napoli L’Orientale
vcaruso@unior.it

Lucia di Pace

Università di Napoli L’Orientale
ldipace@unior.it

Janet DeCesaris and Mercè Lorente Casafont

OLD WORDS, NEW TERMS

Semantic Broadening and Narrowing in the Vocabulary of the Circular Economy

Abstract Although the ‘Circular Economy’ has been widely discussed in the media for years, general dictionaries still do not provide the relevant definitions and/or collocations. We show by examining dictionary definitions that many salient words used in this field have undergone varying degrees of semantic broadening in the general language. Current terminological needs often dictate more precise meanings than those used in the general language, in essence leading to semantic narrowing. We discuss these two seemingly opposing forces in relation to a small set of words from the ‘Circular Economy’ and show how this sort of semantic development is easily accounted for within the Communicative Theory of Terminology.

Keywords Circular Economy; Communicative Theory of Terminology; sense broadening; sense narrowing of terms

1. Introduction

Working towards a sustainable environment is a topic of considerable public interest and as a result, economists, environmental scientists, policy makers and the public at large increasingly refer to the concept of ‘Circular Economy’. As is the case with several sub-fields within a social science, the terminology employed when speaking about this topic is primarily taken from words that already occur in the general lexicon. This type of situation often poses a challenge for defining the terms used in the field, because the expressions being used already exist and often show a significant degree of polysemy and variation in use and meaning.

In this paper, we discuss the development of several salient expressions associated with the terminology of the ‘Circular Economy’ in English and Spanish. The paper is structured as follows. We first briefly set forth the theory of terminology we adopt, the Communicative Theory of Terminology (Cabr , 1999; Cabr , 2023). We then discuss the definition of ‘Circular Economy’ and identify a small number of expressions commonly found in texts on this topic which will be the object of analysis. The fact that several of the lexical items studied have long been in use in general, non-specialized domains has historically led to a significant degree of polysemy, while the recent terminological use of the same expressions leads to semantic narrowing. We also discuss the inclusion and representation of these expressions in current dictionaries of reference of English and Spanish, and conclude by commenting on the terminological needs for this growing field of interest.

2. Communicative Theory of Terminology

The Communicative Theory of Terminology (Cabr , 1999; Cabr , 2023) is an integral approach to terminology that brings together cognitive, social and linguistic factors. Importantly, terminological units are lexical items or lexicalized units, the specialized domain character of which is activated in certain communicative contexts. Terminological units are typically nouns; nevertheless, terminological units may also belong to other lexical categories, such as verbs and adjectives. In the Communicative Theory of Terminology terminological units are linguistic units and as such may display variation in form, meaning and function, and as such the vocabularies of specialized domains display synonymy and polysemy. We believe that this theory is particularly well-suited to analysing terms from specialized domains in which the vocabulary has been taken from existing expressions, as lexical items that have been long been in have often undergone varying degrees of semantic change.

A consequence of Cabr 's linguistic-based approach to terminology is that lexicographic representation in general language dictionaries is not fundamentally different from terminographic representation in special subject domain terminologies because there is no clear-cut structural division between the general lexicon and terms; rather, the difference between lexicographic representation and terminographic representation is one of the thematic constraints on terminographic products. Terminological units, then, may be seen as senses of lexical units that are activated in a context of specialized communication (whether that be between experts or between experts and laymen). In Cabr 's theory, diachronic studies, specifically those on lexical change in special subject domains, are of interest (Lorente Casafont, 2024, pp. 278–279).

It is important to bear in mind that a traditional terminological definition formulated within the General Theory of Terminology (W ster, 1968) typically applies to nouns in the natural sciences and in technological fields and assumes that concepts are easily delimited. W ster and his school of thought believe that concepts can be univocally defined in individual specialized subject domains, which often results in several homonymous terms belonging to different, well-defined subject domains. The Communicative Theory of Terminology differs from the General Theory of Terminology in many aspects that are discussed at length in Cabr  (1999), Cabr  (2003) and elsewhere (see, for example, Faber and L'Homme, 2022) and as such we will not engage in a comparison here, but rather affirm that we have chosen to consider the vocabulary of the 'Circular Economy' through the lens of the Communicative Theory. Given the Communicative Theory's linguistic foundations, a look at the semantic history and use of the main words associated with this topic, as well as their morphological structure, should provide insight into the nature of common words that take on lexicalized, terminological status in certain contexts.

3. The ‘Circular Economy’

3.1 Origin, Definitions, and Inclusion in Dictionaries

What is a ‘Circular Economy’, other than an economy that is in some way opposed to a ‘Linear Economy’? Murray et al. (2017, p. 371) state that it is an economy that has “no net effect on the environment; rather it restores any damage done in resource acquisition, while ensuring little waste is generated throughout the production process and in the life history of the product.” They also state that the origin of the term is debatable, although they place it at an undetermined moment in the 1960s.

The definition given by Murray et al. is not exactly the same as that given by the European Parliament (2016), which states “Circular economy: an economic model based *inter alia* on sharing, leasing, reuse, repair, refurbishment and recycling, in an (almost) closed loop, which aims to retain the highest utility and value of products, components and materials at all times.” We note that the European Parliament’s definition makes no mention of restoring damage to the environment; rather, it concentrates on making optimal use of products (which are shared, leased, reused, repaired, refurbished and recycled).

Neither of these definitions is particularly good by lexicographic standards. In Murray et al.’s definition, “no net effect on the environment” is equated with generating “little waste”, but those two concepts are not equivalent because, presumably, any amount of waste generated has some effect on the environment. The European Parliament’s definition is not much better, as it introduces a hedge (“(almost) closed”) and states that the ‘Circular Economy’ is an economic model based on six different types of actions applied to products and on some other, unidentified actions as well (the reference to “*inter alia*”). Although neither of these works is a dictionary and thus definitions do not have to meet the same standards as those written by lexicographers, they nevertheless introduce elements of ambiguity that could have been avoided, particularly since several of the concepts mentioned are increasingly the object of binding legislation in the European Union (e.g., the right to repair directive adopted by the European Parliament in April 2024, legislation on waste management, legislation on plastic recycling, to name only three (European Parliament, 2024)). In fact, it is precisely because issues associated with the ‘Circular Economy’ are becoming law in an international context that they demand clear, unambiguous and perhaps standardized definitions.

Several—but not all—major dictionaries of English include ‘Circular Economy’ (not necessarily capitalized) as a headword. The *Oxford English Dictionary* currently defines ‘circular economy’ as follows:

An economic system in which the journey of a product, material, etc., leads back in some way to where it began; (now *esp.*) a system or process which seeks to minimize or remediate harm to the environment by recycling, reusing, or regenerating products or materials, as a means of reducing waste and more sustainably or efficiently continuing production.¹

¹ This definition was first published in the OED in 2021.

The first recorded use of this expression dates back to 1932. A Google search for ‘circular economy’ yields a simpler definition from Oxford Dictionaries: “an economic system based on the reuse and regeneration of materials or products, especially as a means of continuing production in a sustainable or environmentally friendly way.” ‘Circular economy’ is defined in the *Collins English Dictionary* (“an economic system based on using renewable resources, eliminating waste, and reusing and recycling material goods”), but is not listed as a lexicalized phrase by *Merriam-Webster Online* in either the *Unabridged Dictionary* or the *Collegiate Dictionary*, nor is it included in either the *American Heritage Dictionary of the English Language* or in *Cambridge English Dictionary Online* (at the date of writing, June 2024). This absence may be due to the fact that the phrase was simply not frequent enough to warrant dictionary inclusion when the dictionary in question was being prepared, or to the fact the lexicalized phrases—as opposed to individual words—are poorly represented in general language dictionaries. In this respect, its absence from the *Cambridge English Dictionary Online* is perhaps surprising because this online resource provides extensive lists of related words and phrases, and includes a vocabulary cloud titled “Environmental issues” containing phrases such as *carbon tax* and *eco-footprint*, both of which are clearly more recent in creation than ‘circular economy’ (the *Oxford English Dictionary* dates *carbon tax* to 1979 and *eco-footprint* to 2002).

Dictionaries of Spanish similarly show uneven inclusion of the expression ‘*economía circular*’. The *Diccionario de la lengua española* by the Spanish Royal Academy in its most recent electronic edition² does not include the phrase. The online commercial lexicographic resource *Diccionarios.com*, however, does include the phrase under the headword ‘*economía*’, and its definition is similar to that given by Oxford Dictionaries, although it also specifically contrasts the concept of ‘Circular Economy’ with that of ‘Linear Economy’:

*Modelo económico que persigue producir bienes de modo sostenible, potenciando la reutilización y el reciclado de los productos más allá de su vida útil: en contraste con la economía lineal, basada en la lógica de usar y tirar, la economía circular busca dar siempre una nueva vida a los productos.*³

3.2 Vocabulary of the ‘Circular Economy’

To determine a small set of lexical expressions frequently used in conjunction with the ‘Circular Economy’ that could be analysed within the scope of this paper, three documents of reference were consulted: The European Parliament’s 2016 briefing on the ‘Circular Economy’ in English; the European Commission’s 2020 “Circular Economy Action Plan for a Cleaner and More Competitive Europe,” in both English and Spanish; and, the “Circular Economy Glossary” published online by the Ellen Macarthur Foundation in both English and Spanish. The following expressions were chosen, as they were all found in all three sources, as well as being prominent in the definitions discussed in the previous section:

² We are referring to the seventh online update of the 23rd edition.

³ Our English translation of the original Spanish: ‘An economic model that aims to produce goods in a sustainable fashion, emphasizing the reuse and recycling of products beyond their useful life: in contrast to the linear economy, which is based on the idea of use and disposal, the circular economy always aims to give products a new use.’

Table 1: Words chosen for study

English	Spanish
<i>circular</i> [adjective], <i>circularity</i>	<i>circular</i> [adjective], <i>circularidad</i>
<i>sustainable</i> , especially <i>sustainable development</i>	<i>sostenible</i> , especially <i>desarrollo sostenible</i>
<i>sustainability</i>	<i>sostenibilidad</i>
<i>recycle</i>	<i>reciclar</i>
<i>reuse</i> [verb]	<i>reutilizar</i>
<i>waste</i> [noun]	<i>residuo/desperdicio/desecho</i>

A cursory examination of these words yields that most are polysemous in general discourse and most are morphological derivatives, and our own knowledge of the history of lexis in English and Spanish indicates that many of these words have been in use in a variety of contexts for generations, if not hundreds of years. We now turn to these characteristics in more detail.

3.2.1 Polysemy, Semantic Development, and Morphological Structure of the English Set of Expressions

The English words under study have all been in use for generations, and in some cases, for centuries. The following table provides an overview of the degree of polysemy recorded in two contemporary dictionaries of reference (one originally published in the United States, the other originally published in Britain), along with the date of first use as reported by *Merriam-Webster.com Dictionary*:

Table 2: Overview, English expressions⁴

	Number of senses, Merriam-Webster	Date of first known use, Merriam-Webster	Number of senses, Collins English Dictionary
<i>circular</i>	6	15 th century	5
<i>circularity</i>	none; run-on entry	not indicated	none; run-on entry
<i>sustainable</i>	2	1924	3
<i>sustainability</i>	none; run-on entry	1924	2
<i>reuse</i>	1	1843	1
<i>recycle</i>	8	1925	4
<i>waste</i>	4	13 th century	10

We note that the differences in the number of senses recorded reflects each dictionary's criteria for determining senses and sub-senses, and a fewer number of senses does not necessarily mean that the dictionary's coverage of meaning is lacking. The fact that all but one of the expressions is polysemous is interesting and reflects the progressive semantic broadening that these expressions have experienced over time.

⁴ The number of senses refers to the major, numbered sense distinctions under the headword and does not include the number of sub-senses.

The original meaning of adjective *circular* clearly refers to round shape and is still frequently used with that sense according to Sketch Engine, but notably the third most frequent noun it modifies in the English Web 2021 corpus is *economy*.⁵ The verb *reuse* has only one sense in both dictionaries, but this is somewhat misleading because the verb on which it is based, *use*, has several current senses in both dictionaries (8 in *Merriam-Webster.com Dictionary* and 6 in the *Collins English Dictionary*).

The oldest English word in this group has undergone semantic broadening over time. The original sense of *waste* is that of ‘barren land’, although today it is more frequently associated with the sense of ‘unwanted by-product of a manufacturing process, chemical laboratory or nuclear reactor’.⁶ According to Sketch Engine’s Word Sketch feature, the five most frequent modifiers of *waste* are *hazardous*, *radioactive*, *toxic*, *solid* and *plastic*, all of which fall under this latter sense and clearly point to the idea of having a harmful effect on the environment.

All of the words with the exception of *waste* are morphologically derived, and the prefix and suffixes present in these words are frequent and still productive in English. The influence of morphological structure on dictionary representation is a complex issue (DeCesaris, 2020–2021) that we will not delve into here, but it is important to note that derived words often display fewer senses than their base forms. This can be observed in the definitions given in *Merriam-Webster.com Dictionary* for *sustain* and *sustainable*, as the transitive verb *sustain* has 8 senses, most of which are not applicable to the use of the adjective *sustainable* nor to its derived noun, *sustainability*.

Table 3: Definitions of *sustain* and *sustainable*, Merriam-Webster.com Dictionary

	<i>sustain</i>	<i>sustainable</i>
Sense		
1	to give support or relief to	capable of being sustained
2	to supply with sustenance : NOURISH	a: of, relating to, or being a method of harvesting or using a resource so that the resource is not depleted or permanently damaged <i>sustainable techniques</i> <i>sustainable agriculture</i> b: of or relating to a lifestyle involving the use of sustainable methods <i>sustainable society</i>
3	KEEP UP, PROLONG	
4	to support the weight of : PROP <i>also</i> : to carry or withstand (a weight or pressure)	
5	: to buoy up <i>sustained</i> by hope	

⁵ The most frequent noun modified by *circular* is *motion*, followed by *saw*, *economy*, and *orbit*.

⁶ Definitions have been adapted from *Merriam-Webster.com Dictionary*.

6	a: to bear up under b: SUFFER, UNDERGO <i>sustained</i> heavy losses	
7	a: to support as true, legal, or just b: to allow or admit as valid the court <i>sustained</i> the motion	
8	to support by adequate proof: CONFIRM testimony that <i>sustains</i> our contention	

It is not uncommon in dictionaries of English for morphologically derived words to remain undefined if the dictionary in question defines both the base word and the pertinent affix as a headword (which is the case for *circularity* and *sustainability* in *Merriam-Webster.com Dictionary*; *-ity* is defined as a noun suffix meaning “quality, state, degree”).

Examination of usage data confirms that *sustain* occurs in a wide variety of contexts, whereas *sustainable* currently refers to economics. The data from Sketch Engine (English Web 2021 corpus) shown in Table 4 for the use of *sustain* as opposed to the use of *sustainable* reveal that what is *sustained* in English is quite varied, ranging from *injuries*, *damage*, *wounds*, and *loss* to *growth* and *momentum*. We note that what is *sustained* may have either positive semantic prosody (*growth*, *momentum*) or negative semantic prosody (*injuries*, *damage*, *wounds*, *loss*), and the direct object of the active verb can be either (cf. *sustain damage* vs. *sustain momentum*), although the pattern [**noun** *sustained*] with the postposed participle strongly associates with negative prosody. In contrast, *sustainable* is mainly used to modify *development*, *future*, *goal*, *agriculture*, and *growth*, all of which are part of the ‘Circular Economy’.

Table 4: Comparison of Word Sketch data for *sustain* and *sustainable*

Objects of <i>sustain</i>	Nouns modified by <i>sustainable</i>
injury injuries sustained	development sustainable development
damage sustained damage	future a sustainable future
wound wounds sustained	goal sustainable development goals
loss losses sustained	agriculture sustainable agriculture
growth sustain growth	growth sustainable growth
momentum sustain the momentum	practice sustainable practices
fracture sustained a fracture	management sustainable management
concussion sustained a concussion	livelihood sustainable livelihoods

3.2.2 Polysemy, Semantic Development, and Morphological Structure of the Spanish Set of Expressions

The Spanish words under consideration all have Latin roots, and all but *residuo* are morphologically derived.⁷ The degree of polysemy represented in dictionaries is somewhat less than that observed in English, as shown in Table 5.⁸

Table 5: Overview, Spanish expressions

	Number of senses, <i>Diccionario de la lengua española, Real Academia Española</i>	Number of senses, <i>Diccionarios.com</i>	Comparison with number of English senses, <i>M-W.com / Collins</i>
<i>circular</i>	4	4	6 / 5
<i>circularidad</i>	1	0	0 / 0
<i>sostenible</i>	2	2	2 / 3
<i>sostenibilidad</i>	1	1	0 / 2
<i>reciclar</i>	4	3	8 / 4
<i>reutilizar</i>	1	1	1 / 1
<i>residuo</i>	4	3	
<i>desperdicio</i>	2	2	
<i>desecho</i>	5	3	

We note that the derived words *circularidad*, *sostenibilidad*, and *reutilizar* are defined in the dictionaries with a clear cross-reference to their base words, which is common practice in Spanish dictionaries (Spanish dictionaries do not have a strong tradition of run-on entries, unlike English dictionaries, and they often require the user to refer to definitions under another headword). The verbs with the prefix *re-* are relatively new additions to the Academy dictionary (*reciclar*, in 1984; *reutilizar*, in 2014).

Interestingly, texts on the ‘Circular Economy’ in Spanish use three equivalents for the word *waste* in English: *desecho*, *desperdicio*, and *residuo*. All three have been in use in Spanish for hundreds of years, having been recorded in the earliest dictionaries of the language. The current definitions of these words do not clearly distinguish among the three; for example, the Academy dictionary gives *residuo* as the third sense of *desecho*, and states that both *residuo* and *desperdicio* are synonyms of *desecho*.

⁷ The word *residuo* was borrowed from the Latin into Spanish in the 15th century and cannot be considered a morphologically derived word in a synchronic analysis of Spanish.

⁸ No comparison is shown for *residuo/desperdicio/desecho* because all correspond to *waste*.

Table 6: Definitions, *Diccionario de la lengua española*, 23rd edition⁹

<i>desecho</i>	<i>desperdicio</i>	<i>residuo</i>
1. m. Aquello que queda después de haber escogido lo mejor y más útil de algo. SIN.: restos, sobras, <u>residuo</u> , piltrafas.	1. m. Derroche de la hacienda o de otra cosa. SIN.: despilfarro, derroche, dilapidación. ANT.: aprovechamiento	1. m. Parte o porción que queda de un todo. SIN.: resto, sobrante, excedente, remanente, saldo, resta, rezado, pucho resquicio.
2. m. Cosa que, por usada o por cualquier otra razón, no sirve a la persona para quien se hizo.	2. m. Residuo de lo que no se puede o no es fácil aprovechar o se deja de utilizar por descuido. SIN.: restos, broza, <u>desecho</u> , <u>residuo</u> , despojos. ANT.: aprovechamiento	2. m. Aquello que resulta de la descomposición o destrucción de algo.
3. m. Residuo, basura. SIN.: basura, <u>desperdicio</u> , escoria, <u>residuo</u> , bazofia, broza, piltrafas.		3. m. Material que queda como inservible después de haber realizado un trabajo u operación. U. m. en pl. SIN.: restos, sobras, despojos, <u>desecho</u> , <u>desperdicio</u> , basura, escombros, morralla, escurriduras.
4. m. Desprecio, vilipendio.		4. m. Mat. Resto de la sustracción y de la división. SIN.: resto, diferencia, resta, restante.
5. m. Lo más vil y despreciable.		

These three words have been defined in similar terms for hundreds of years. For example, the definition for *desecho* in the 1780 edition of the Academy dictionary reads as follows: “*Sobra, desperdicio de alguna cosa, el residuo que queda despues de haber escogido lo mejor y util de ella.*”¹⁰

Corpus data do not clarify the issue much, as all three words are commonly used as direct objects of the verbs *expulsar* ‘expel’, *reutilizar*, *verter* ‘dump’, and *reciclar*. European Union documents tend to refer to *residuos* in Spanish, but we note that the concept of *residual waste* is usually rendered as *desechos residuales*. Although *desperdicio* is common in texts with reference to food, it is not exclusively used as such: although the most frequent modifier of *desperdicio* is *alimentario* (‘food-related’), the second most frequent is *industrializable* (‘able to be industrialized’).¹¹ It often implies that the waste was not a necessary outcome of the process involved. We have not been able to discern a clear-cut difference between *desecho* and *residuo*, and frequency data from Sketch Engine similarly show that both are in current use and collocate with the same words (*sólido* ‘solid’, *tóxico* ‘toxic’, *peligroso* ‘dangerous’, *radioactivo* ‘radioactive’, among others).

⁹ Explanation of abbreviations used in the *Diccionario de la lengua española*: *m.* (masculine gender for a noun); *SIN* (synonyms); *ANT* (antonyms); *U. m. en pl.* (mainly used in the plural); *MAT* (mathematics). Senses that are restricted to a specific geographical area and which are so labelled have not been included in this table. The underlining in red is ours.

¹⁰ The 1780 edition was consulted via the *Nuevo Tesoro Lexicográfico de la Lengua Española*. Our English translation of the original Spanish (which was printed without accents): ‘Leftover, waste from something, the waste that remains after having chosen the best and most useful parts of it.’

¹¹ According to Word Sketch, Sketch Engine (Spanish Web Corpus 2018).

4. Discussion

The lexical items most closely associated with the ‘Circular Economy’ are all in general use in both English and Spanish in a variety of semantic contexts. The key word *circular* may have its origins in a description of a physical reality, but clearly over time its use has been extended to intangible contexts: shapes, motions, orbits, arguments, and economies can all be *circular*. The fact that we now use the lexicalized phrase ‘Circular Economy’ with a fixed, terminological-like meaning does not preclude the use of *circular* in other contexts, some of which display differing pragmatic overtones: *circular* is now a positive characteristic of manufacturing processes, but is a negative feature of arguments or reasoning. Similar observations may be made about other lexical items in this field. *Sustain*, and especially its participial form *sustained*, is commonly used in court proceedings in English (see sense 7 in Table 3 above), but *sustainable* increasingly—and almost exclusively—is used to refer to initiatives that do not harm the environment. The meanings of *sustain/sustainable/sustainability*, like those of the Spanish equivalents *sostener/sostenible/sostenibilidad*, have expanded over time, but in the more recent usage in specialized discourse on waste management and environmental policy, the meanings of these words have become narrower. In essence, the use of these words in a terminological context has initiated the reverse of the historical trend of semantic broadening.

The use of old words from the general language in a new specialized field, in which words are used with terminological intent, is nicely accounted for by the Communicative Theory of Terminology. The terminological sense of expressions is triggered by their use in specialized discourse; e.g., in texts on economics, *sustainable* in the expressions *sustainable development* or *sustainable future* makes use of a specific meaning of *sustainable*. The definition of the terminological sense may be phrased following the same style requirements adopted by a specific dictionary project and does not need to be structurally different from the style used for defining other senses. In our view, then, it is not that *circular* in physics, for example, is a different term from *circular* in economics; rather, the word *circular* is polysemous and takes on different nuances of meaning when used in domain-specific discourse. Those nuances are typically represented in dictionaries as separate senses, often with subject labels.

Given the current social interest in the ‘Circular Economy’, it is somewhat disheartening that several of the dictionaries examined did not afford the expression headword status. We note that the newer sense of *sustainable/sostenible* is now widely accounted for in general language dictionaries, and we would expect that ‘Circular Economy’ will soon find its way into more headword lists. A good dictionary definition of ‘Circular Economy’ would improve on those given in section 3.1 of this paper by eliminating the ambiguity those definitions display; the European Parliament’s definition in particular could be improved by specifying what ‘*inter alia*’ refers to, and this may require substantial rewording of the definition.

5. Concluding Remarks

The lexis in the 'Circular Economy' provides for a fascinating case of rapidly developing senses for words that have existed for a long time. We have showcased the tension between general lexicological development and specific terminological needs. The words most associated with this field are polysemous, yet present terminological requirements that select and prioritize specific senses in specialized discourse. The need for improved lexicographic representation of the vocabulary in this field is clear, as governments worldwide increasingly enact legislation on the 'Circular Economy' that is important to us all. If that legislation is to be binding and enforceable by law, the expressions used must be clearly defined.

References

- Cabré, M. T. (1999). *La terminología. Representación y comunicación*. Institut Universitari de Lingüística Aplicada.
- Cabré, M. T. (2003). Theories of terminology: Their description, prescription and explanation. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 9(2), 163–199. <https://doi.org/10.1075/term.9.2.03cab>
- Cabré, M. T. (2023). *Terminology. Cognition, language and communication*. John Benjamins.
- Circular Economy Glossary / Glosario de economía circular. (2021). Ellen Macarthur Foundation. Retrieved February 4, 2024, from <https://ellenmacarthurfoundation.org>
- DeCesaris, J. (2020–2021). Dictionaries and Morphology. In Z. Gavriilidou, L. Mitits, & S. Kiosses (Eds.), *Lexicography for Inclusion: Proceedings of the 19th EURALEX International Congress*, Vol. 2. (pp. 577–584). Democritus University of Thrace.
- European Commission. (2020). *A new Circular Economy Action Plan for a cleaner and more competitive Europe / Nuevo Plan de acción para la economía circular por una Europa más limpia y más competitiva*. Document 52020DC0098. Retrieved January 22, 2024, from <https://eur-lex.europa.eu>
- European Parliament. (2016). Briefing. Closing the Loop. New Circular Economic Package. Document 5738999. Retrieved January 15, 2024, from [https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI\(2016\)573899](https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI(2016)573899)
- European Parliament. (2024). Topics. Climate and environment. Circular economy. Retrieved July 25, 2024, from <https://www.europarl.europa.eu/topics/en/topic/circular-economy>
- Faber, P., & L'Homme, M.-C. (2022). *Theoretical Perspectives on Terminology. Explaining terms, concepts and specialized knowledge*. John Benjamins.
- Lorente Casafont, M. (2024). La lexicalització dins del canvi terminològic. *Caplletra. Revista Internacional De Filologia*, 76, 277–298. DOI: <https://doi.org/10.7203/caplletra.76.28217>

Murray, A., Skene, K., & Haynes, K. (2017). The Circular Economy: An Interdisciplinary Exploration of the Concept and Application in a Global Context. *Journal of Business Ethics* 140, 369–380. <https://doi.org/10.1007/s10551-015-2693-2>

Sketch Engine. Word Sketch [sustain], [sustainable], [waste], [residuo], [desperdicio], [desecho]. Retrieved May 24, 2024, from <https://www.sketchengine.eu/guide/word-sketch-collocations-and-word-combinations/>

Wüster, E. (1968). *The Machine Tool. An interlingual dictionary of basic concepts*. Technical Press.

Dictionaries:

American Heritage Dictionary of the English Language Online. <https://www.ahdictionary.com/>

Cambridge English Dictionary Online. <https://dictionary.cambridge.org/dictionary/>

Collins English Dictionary Online. <https://www.collinsdictionary.com/dictionary/english>

Diccionario de la lengua española. 23rd edition, 7th online update. Real Academia Española. <https://dle.rae.es>

Diccionarios.com. <https://www.diccionarios.com>. [Portal of Larousse Editorial in Spain and compendium of dictionaries edited by Larousse and Vox]

Merriam-Webster.com Dictionary. <https://www.merriam-webster.com/dictionary>

Nuevo Tesoro Lexicográfico de la Lengua Española. <https://apps.rae.es/ntlle/SrvltGUILoginNtlle>

Oxford English Dictionary Online. <https://www.oed.com>

Acknowledgements

Work on this paper was supported by the project EVOTERM. *Procesos de lexicalización y complejidad de las comunidades de hablantes en el estudio diacrónico de la terminología y del discurso de especialidad en economía e informática*, reference PID2021-127585NB-I00, financed by the *Agencia Estatal de Investigación* (AEI) of Spain and the European Regional Development Fund (ERDF).

Contact information

Janet DeCesaris

Universitat Pompeu Fabra (retired)
janet.decesaris@upf.edu

Mercè Lorente Casafont

Universitat Pompeu Fabra / Institut d'Estudis Catalans
merce.lorente@upf.edu

Maucha Gamonal, Adriana Pagano, Tiago Torrent, Ely Matos,
and Arthur Lorenzi

AUTOMATED SEMANTIC FRAME ANNOTATION

An Exploratory Study in the Health Domain

Abstract A medication package insert is a legal healthcare document with important information about medications. In Brazil, the National Health Surveillance Agency (ANVISA) requires two versions of the package insert: one for patients and another one for healthcare professionals. In this study, we manually evaluated the performance of an automatic frame annotator on a corpus consisting of 100 sentences targeting patients and 100 sentences targeting healthcare professionals. The aim was to evaluate whether the parser's output evidenced correct assignment of semantic frames and their frame elements (FEs) in each input sentence and to what extent human post-annotation would be necessary to improve the output. Text target audience was defined as a variable potentially impacting frame detection given differences in the language of package inserts. Overall, the findings demonstrate the efficacy of the automated annotation process, revealing challenges that have to do with the same form being capable of classification under different categories and/or frames. Few differences were found when comparing sentences for the two distinct audiences targeted in the texts.

Keywords FrameNet; FrameNet Brasil; Large Ontology Multilingual Extraction (LOME); FrameNet parser; automated semantic frame annotation; manual annotation; health domain

1. Introduction

The theory of Frame Semantics (Fillmore, 1982) relates lexical meaning to human experience, whether perceptual, social, or cultural. Semantic frames are structures which organize such knowledge. Through FrameNet, we can relate context and meaning in a multidimensional and multimodal approach (Torrent et al., 2022).

FrameNet, which originated with an interest in computational lexicography, has now expanded to various uses in language technology (Belcavello et al., 2024; Viridiano et al., 2024). In this work, we present initial contributions on the results of a semantic parser powered by FrameNet's linguistic annotation data (Das et al., 2014).

To this end, we explored the use of a semantic frame parser and human evaluation of its output in an annotation effort towards developing resources for Brazilian Portuguese. Our aim was to assess whether the parser's output showed correct assignment of semantic frames and their frame elements (FEs) in each input sentence and in cases of incorrect assignment what type of human post-annotation would be necessary to improve the output. Additionally, we wanted to find out if the parsing output would be impacted by the register of the texts used.

In this annotation task, a small corpus of 200 sentences was compiled from the dosage and administration section in medication package inserts written in Brazilian Portuguese and published by Brazil's National Health Authority ANVISA. Our study focused on texts targeting two distinct audiences, with 100 sentences addressing lay users/patients and 100 sentences addressing healthcare professionals. Target audience was defined as a variable potentially impacting frame detection as the parser is modelled upon language input which does not necessarily include samples of text with the characteristics of the language used to address each targeted group.

The sampled sentences were automatically annotated using the semantic frame parser in the LOME system (Xia et al., 2021), a multilingual information extraction system, fine-tuned with a corpus of annotated texts in Brazilian Portuguese. The texts are part of FrameNet Brasil database, which includes a model as well as texts annotated for the health domain (Dutra et al., 2023).

The paper is organized as follows: Section 2 presents the semantic annotation process of FrameNet. In this section, we outline the key concepts of the FrameNet model and describe the semantic annotation process within this framework. Section 3 describes the annotation task, providing information on the compiled corpus, the parser used, the manual annotation stages, and the types of post-editing performed. In Section 4, we discuss the annotation and post-editing results, illustrating our findings with examples. Finally, Section 5 concludes the paper and outlines future projections for this work.

2. FrameNet Semantic Annotation

2.2 FrameNet Model

FrameNet is a semantically oriented computational resource drawing on Frame Semantics (Fillmore, 1982; 1985) for the following concepts: frame, frame element, lexical units, and frame-to-frame relations.

Frame: a frame is a linguistic-computational system used for knowledge representation, modelling concepts and situations across different domains. For instance, the `waking_up`¹ frame describes the transition from a state of consciousness where an individual is largely unaware of their environment to a wakeful state.

Frame Element (FE): a frame consists of elements, so-called frame elements. They are participants, props, and other fundamental elements that define the frame. Within its structure, FEs are categorized into nuclear and non-nuclear elements. In the `waking_up` frame, the nuclear FEs are `SLEEP_STATE` and `SLEEPER`, which are essential for conceptualizing the sleeping event. Other FEs, such as `TIME`, `PLACE`, and `CIRCUMSTANCES`, can be instantiated within the scope of a sentence, though not necessarily so, and are hence considered non-nuclear.

¹ Following established conventions, frame names are set in `Courier`, frame elements in `SMALL CAPS`.

In the process of semantic annotation, FrameNet identifies instances of null instantiation of FEs, namely Definite Null Instantiation (DNI), Indefinite Null Instantiation (INI), and Constructional Null Instantiation (CNI). Briefly stated, a DNI is an example of anaphoric omission, where the missing FE has been previously mentioned in the linguistic or discursive context. An INI comprises situations where the FE cannot be inferred from the context. A CNI is characterized by a non-realization licensed by a grammatical construction, such as a subject not realized by a pronoun in imperative sentences.

Lexical Unit: a lexical unit is the linguistic material that instantiates a frame, i.e., a lemma endowed with a specific meaning (form + meaning). In the sentence, “*This morning, Maria woke up from a long dream*”, we analyse the lexical unit *wake up.v* as evoking the **Waking_up** frame. “*Maria*” occupies the slot of the nuclear FE **SLEEPER**, while “*from a long dream*” occupies the slot of the nuclear FE **SLEEP_STATE**. “*This morning*” occupies the slot of the non-nuclear FE **TIME**.

In Framenet, there are generally two ways to view the semantic annotation of lexical units, as shown by the two representations of the sentence in Example 1 “*This morning Maria woke up from a long dream.*”

Representation 1

[*This morning*_{TIME}], [*Maria*_{SLEEPER}] **WOKE UP** [*from a long dream*_{SLEEP_STATE}]

Representation 2

This morning, *Maria* **WOKE UP** *from a long dream*².

Associations between form and meaning which evoke lexical frames may involve multiword expressions (MWEs). An MWE is the association of two or more lexemes that implicate a unit of meaning and evoke a frame. Some idiomatic expressions are good examples of MWEs, such as ‘middle of nowhere,’ as explained in Ruppenhofer et al. (2016). If we consider the study of specialized domains from a terminological perspective, multi-word units are defined based on the relevance of a given terminological unit to the technical domain in question. In this work, although we are dealing with texts belonging to a specific domain, we have not annotated terminological units; rather, we have focused on lexical LUs and relied on annotations previously stored in our database.

To the basic concepts hitherto presented we should add that of frame to frame relation.

Frame to frame relation: the relationship between frames reinforces the assumption that language knowledge is interconnected and that a frame interacts with other frames. Fillmore & Baker (2009) present a set of frame to frame relations in three major groups: generalization (inheritance, perspective_on, using), event structure (subframe, precedes) and systematic relations (causative_of, inchoative_of).

² Each label has a colour corresponding to a Frame Element instantiated in sentences from corpora. Examples of corpus annotation in English are found in the original FrameNet (framenet.icsi.berkeley.edu). For linguistic data in Brazilian Portuguese, these can be accessed through the webtool annotation platform (webtool.frame.net.br).

Returning to our example, *Waking_up* is a subframe of *Sleep_wake_cycle*, meaning it is a sub-event of a more complex event. *Waking_up* precedes *Being_aware* and is preceded by *Sleep*. Frame to frame relations are detailed in Ruppenhofer et al. (2016).

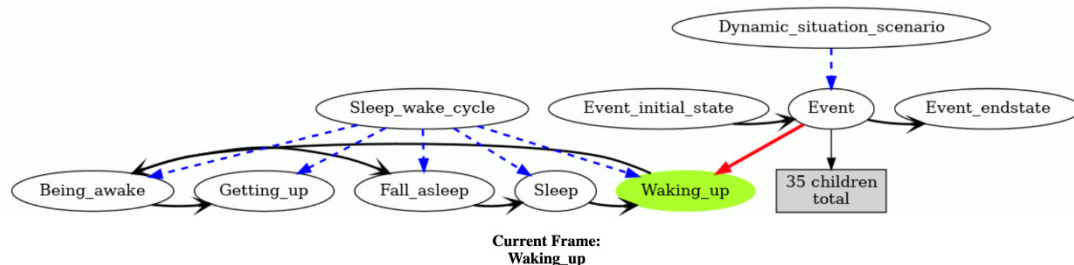


Fig. 1: Waking_up frame to frame relation³

2.3 FrameNet Annotation

FrameNet linguistic annotation records semantic and syntactic information of a Lexical Unit. Depending on the purpose, the annotation can be carried out in two modes: lexicographic or full text annotation. In lexicographic annotation, the annotator annotates various instances of corpora of the same lexical unit. This is the most common form of annotation. The purpose is to evaluate the semantic and syntactic behaviour of the same lexical item from a specific meaning.

The second mode is full text annotation, whereby all the frames evoked within the scope of a sentence are annotated.

In our work, we carried out full text annotation to evaluate the performance of an automatic annotator in assigning frames and their elements in sentences retrieved from text in medication package inserts. Figure 2 shows a full text annotation performed by our frame semantic parser for Example (2), and Figure 3 shows a screenshot of our software interface where sentence (2) was manually revised and edited.

The sentence in Example (2) was retrieved from a patient medication package insert

1. Lave suas mãos após a aplicação do produto
Wash your hands after the application of the product
“Wash your hands after applying the product”

[Lave ^{GROOMING FRAME}] suas [mãos ^{BODY_PART FRAME}] [após ^{TIME_VECTOR FRAME}] a [aplicação ^{PLACING FRAME}] do [produto ^{ENTITY FRAME}].

Note that each Lexical Unit (LU) is identified by its corresponding frame followed by its Part of Speech (POS), for example, Grooming.lave.v. (Grooming = frame; *lave* [*wash*] = word form of the verb *lavar* [*to wash*]; v=verb).

³ <https://framenet.icsi.berkeley.edu/node/5039>

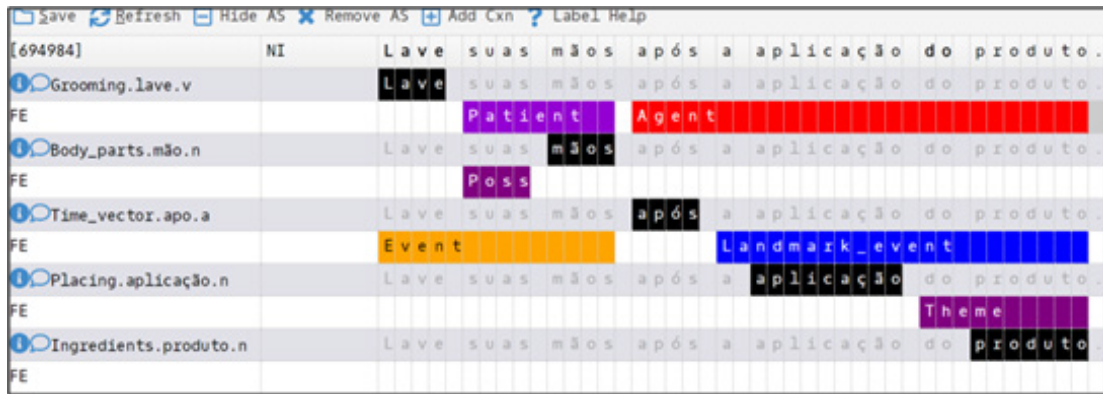


Fig. 2: Screenshot showing full text annotation performed in a frame semantic parser [webtool.framenetbr.uff.br]

The output of the automatic annotation was post-edited for different types of error using the FrameNet interface. For example, post-editing was carried out to fix tokenization errors (*apo.a* instead of *após* [after]) as well as POS tagging errors (verb *dosar* [to dose].v instead of noun *dose* [dose].n). Incorrect frame assignment was also edited. For instance, in Figures 2 and 3, we can see five frames being evoked in each. Except for *produto* [product].n, all other lexical items were correctly assigned to their frames.

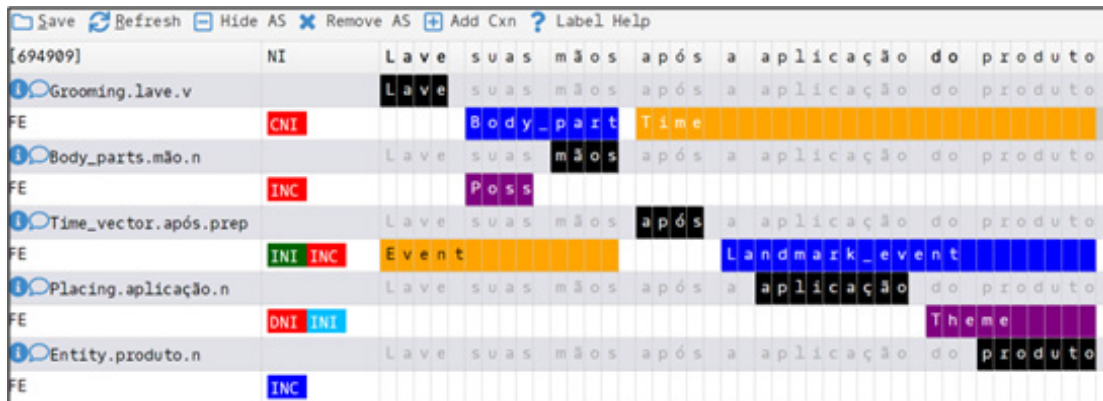


Fig. 3: Full text annotation manually revised and edited [webtool.framenetbr.uff.br]

Regarding the Frame Elements (FEs), they were correctly assigned, except for the FEs of the LU *lavar* [to wash].v, where “*suas mãos*” [your hands] was tagged as PATIENT, while the correct tag should be BODY_PART, and “*após a aplicação do produto*” [after the application of the product] was marked as AGENT, whereas the correct tag is TIME. Table 1 shows the semantic annotation for that sentence. During the review and editing process, manual insertion of null instantiations was performed.

Table 1: Semantic annotation for sentence in Example 2

Lexical Unit	Frame	Frame Element
<i>lavar.v</i>	Grooming	BODY_PART, TIME
<i>mãos.n</i>	Body_part	POSSESSOR
<i>após.prep</i>	Time_vector	EVENT, LANDMARK_EVENT
<i>aplicação.n</i>	Placing	THEME
<i>produto.n</i>	Entity	-

3. The Annotation Task

Given the centrality of developing tools for automating semantic annotation, this work undertook the task of post-editing annotations in order to evaluate types of necessary modifications upon automatic annotation, following FrameNet's semantic annotation criteria.

3.1 Corpus

A medication package insert is a relevant text for obtaining safe information about the safe use of a medication and its potential risks. In Brazil, this text is available in a more technical version aimed at healthcare professionals and a less technical version intended for patients or their caregivers, such as in the case of medication for children.

The text in a medication package insert comprises several sections. Examples of these sections include presentation, composition, contraindications, dosage and administration, among others. Each country has its own regulatory agency. In the United States, the Food and Drug Administration is the agency that sets the guidelines for these documents. In Brazil, the National Health Surveillance Agency (ANVISA) is responsible for controlling such information. Package inserts can be accessed online at: <https://consultas.anvisa.gov.br/#/bulario/>, a website frequently updated by ANVISA.

In our study, we retrieved medication package inserts from ANVISA's website. We chose the dosage section to compile 200 sentences for analysis, with one hundred sentences aimed at healthcare professionals and one hundred at patients. Our corpus totals 3615 words and is sentence-aligned, so that for each sentence collected from the dosage section in the text addressing healthcare professionals there is a counterpart sentence in the text addressing the patient. The task of collecting and aligning the corpus was carried out manually by junior researchers participating in the project.

Unlike studies which conducted a quantitative evaluation of automatic frame parsing performance (Das et al., 2010; Zadeh et al., 2019), our work pursued a qualitative approach and due to the small size of the corpus, a statistical approach that would allow for measuring the parser's performance was not conducted. However, important insights were gathered from the output.

3.2 Automated Semantic Frame Annotation

For the automatic annotation of frames and frame elements in this work, the LOME (Large Ontology Multilingual Extraction) system was used. Its pipeline includes a full FrameNet parser, capable of identifying the frames and frame elements in a sentence (with their occurring spans), as well as performing coreference resolution, entity typing and temporal resolution between events (Xia et al., 2021).

Because it is designed for multilingual information extraction, LOME is trained on top of the XLM-RoBERTa multilingual large language model (Conneau et al., 2020), allowing it to learn in one language and perform semantic parsing in others.

In this work, we used a version of LOME trained with full-text annotations from Berkeley FrameNet 1.7 and full-text annotations from FrameNet Brasil 1.7. The training data from FrameNet Brasil includes a set of recently created frames for the healthcare domain (Dutra et al., 2023) and their related full-text annotations for Brazilian Portuguese.

What distinguishes LOME from other parsers is having only a sentence as input. Other parsers, for example, can only identify a frame if the trigger span is part of the input. A trigger span is a continuous part of the sentence (one or more tokens) that evokes a frame. These are expected to be the spans of lexical units. The main difference is that in human-annotations, the lexical units of a frame are defined and then used to annotate. LOME does not have any kind of information on lexical units, so it just searches for spans that evoke a frame. In a scenario where the model is 100% accurate, every trigger span is the span of a lexical unit.

3.3 Manual Annotation Effort

The LOME results were imported into the FN.Br annotation software for the task of manual post-editing. Figure 4 illustrates the revision workflow. Initially, annotators verified whether the target was correctly identified, followed by checking if the segmentation of the target unit and part-of-speech tagging were correctly done. If not, manual revisions were performed. Subsequently, the assessment of frame assignment was conducted. If the lexical item had been correctly assigned to the frame, the next step was to evaluate whether nuclear and non-nuclear frame elements were identified and consequently labelled.

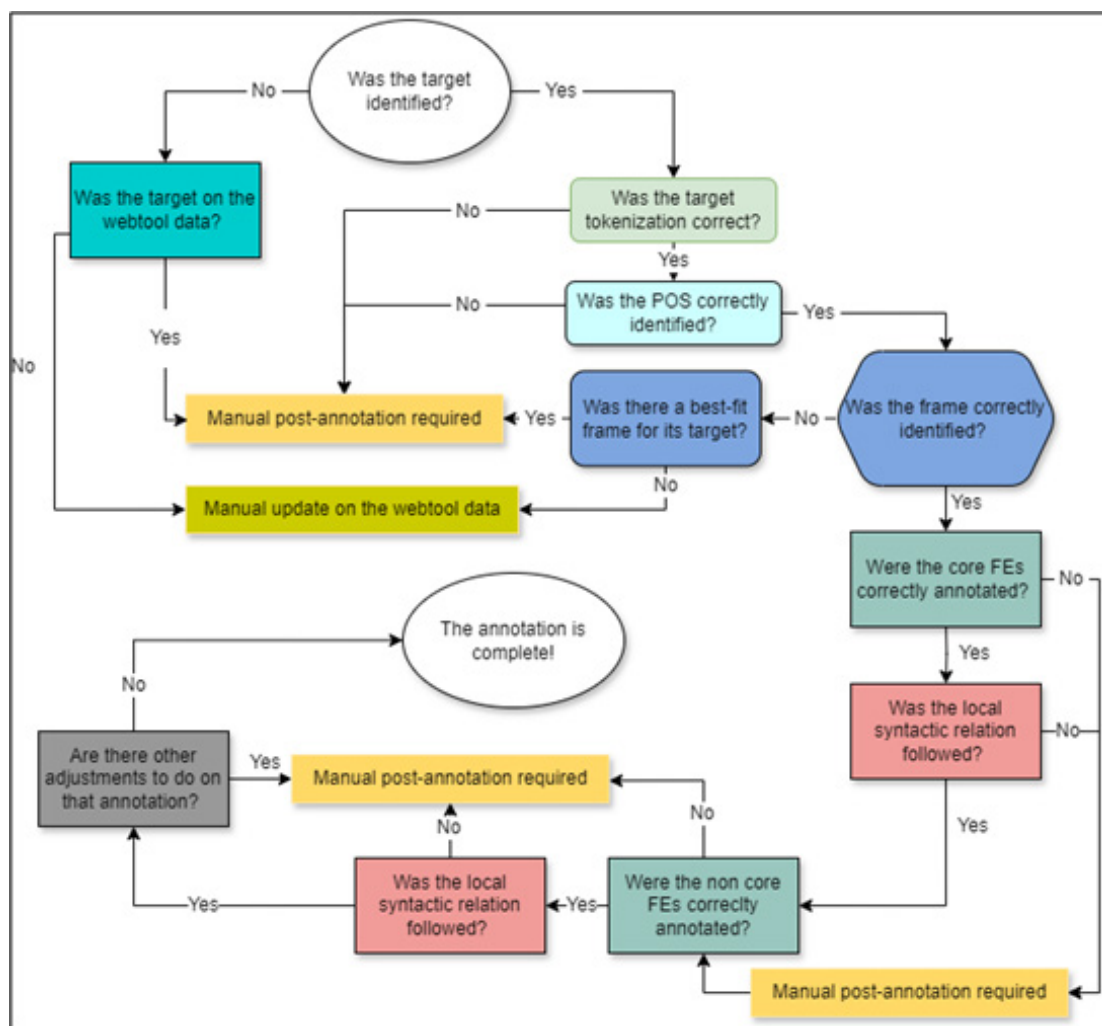


Fig. 4: Post-annotation flowchart

Among the post-annotations made based on the results of the automatic annotation, the main ones are as follows:

- **Target:** target identified or not.

In the sentence in Example 2, the target *dever.v* was the only one recognized as a target by LOME, and the **Desirable_event** frame was assigned to the target. That sentence was post-annotated and Example 3 shows the other targets that evoke frames.

- JANUMET **deve** ser administrado duas vezes ao dia, durante as refeições.
JANUMET should be administered twice a day, during meals
'JANUMET should be administered twice a day, during meals.'
- JANUMET **deve ser administrado duas vezes ao dia, durante as refeições.**

- **Tokenization:** correctly identified or not.

Errors in target tokenization were identified in several sentences, as shown in Figure 2 and illustrated in Example 4 with the target *após*.prep.

4. Não utilize nenhuma bandagem (faixa) **após** aplicação do medicamento.
Do not use any bandage after applying the medication
'Do not use any bandage after applying the medication.'

- **Part-of-speech:** correctly identified or not.

The LOME system was not trained for Part of Speech (POS) recognition. In many instances of correct semantic frame assignment, manual intervention was required to properly assign the POS of the word under analysis. In Example 6, *uso.n* (*use.n*) was identified as *usar.v* (*use.v*) in the **Using** frame. Through manual editing the POS tag was changed to *uso.n* in **Using** frame.

5. Evitar o **uso** de roupas íntimas de tecido sintético (como nylon), utilizar as de algodão.
Avoid the use of underwear synthetic fabric (such as nylon) prefer those of cotton
'Avoid the use of synthetic fabric underwear (such as nylon), prefer cotton ones.'

- **Frame:** correctly inferred or not.

In Example 6, the best-fit frame is **Change_position_on_a_scale**, as this frame indicates the change in position of an item on a scale from an initial point to an end point. The semantic frame parser directed the target to the **Health_condition** frame. However, *depressão.n* (*depression.n*) in the **Health_condition** frame refers to a psychiatric disorder, which is not the case in Example 6.

6. Os pacientes devem ser bem monitorados quanto ao desenvolvimento de **depressão** profunda da medula óssea.
The patients should be closely monitored for the development of depression deep marrow bone
'The patients should be closely monitored for the development of deep bone marrow depression.'

- **Frame Element (core or non-core):** addition, suppression, modification of FE.

Example 7 illustrates the identification of the frame and its FEs by manual post-annotation and by the semantic frame parser. The lexical unit is *usar.v* and the frame is **Using**. This frame involves an agent manipulating an instrument to achieve some purpose. The core FEs are AGENT, INSTRUMENT, and PURPOSE. Among the non-core FEs are PLACE, TIME, FREQUENCY, and several others.

7. Semantic parser annotation:

Use ^{USAR.V: USING FRAME} [o aplicador _{INSTRUMENT}] [apenas uma vez ^{TEMPO FREQUENCY}].

Manual post-annotation:

Use ^{USAR.V: USING FRAME} [o aplicador _{INSTRUMENT}] [apenas uma vez _{FREQUENCY}].

Use the applicator only once

‘Use the applicator only once.’

The AGENT and PURPOSE are not instantiated in the sentence. In this case, the AGENT is marked in the manual annotation as CNI due to the imperative. The INSTRUMENT FE is instantiated and was correctly identified as “o instrumento” (the instrument). On the other hand, what the parser identified as TIME is the FE FREQUENCY, “apenas uma vez” (only once). Based on the types of edits performed, this example includes a FE modification.

4. Results

A series of errors were identified regarding the assignment of frames and FEs. Among them, polysemous targets stood out. In the sentence presented in Table 2, there is a warning for patients in case they forget to take their medication: they do not need to take the missed dose, but should follow the usual schedule for the next dose.

Table 2: The polysemy of *esquecido.a* (missed.a)

PATIENT	HEALTHCARE PROFESSIONAL
Caso você se esqueça de tomar rosuvastatina cálcica, não é necessário [tomar a dose ^{ACTION}] esquecida .	Se o paciente se esquecer de tomar uma dose de rosuvastatina cálcica, não é necessário tomar a dose esquecida, deve-se apenas tomar a próxima dose, no horário habitual.
‘If you forget to take rosuvastatin calcium, it is not necessary to take the missed dose.’	‘If the patient forgets to take a dose of rosuvastatin calcium, it is not necessary to take the missed dose; simply take the next dose at the usual time.’

Considering FN.BR database, the *esquecido.a* lemma can be associated with different conceptual frames, such as Remembering_information, Remembering_to_do, Remembering_experience, and Abandonment. The frame semantic parser assigned *esquecida.a* to the Abandonment frame. In this frame, an AGENT leaves behind a THEME effectively rendering it no longer within their control or of the normal security as one’s property. However, Remembering_to_do seems to be the best-fit frame, since a COGNIZER thinks of and performs an ACTION that is a self- or other-imposed task or some other kind of desirable behavior. The action may involve a SALIENT_ENTITY in some way affected by the cognizer. If a SALIENT_ENTITY is mentioned, the ACTION is left unexpressed.

Table 3: The polysemy of *dever.v* (*should.v*)

PATIENT	HEALTHCARE PROFESSIONAL
No caso de se esquecer de uma dose de OTEZLA, tome-a assim que se lembrar. 'If you miss a dose of Otezla, take it as soon as you remember.'	Se o paciente esquecer uma dose, [a próxima dose ^{REQUIRED_SITUATION}] deve [ser tomada ^{REQUIRED_SITUATION}] [o quanto antes ^{TIME}]. 'If the patient misses a dose, the next dose should be taken as soon as possible.'
Em bebês e crianças de até 4 anos de idade, [o produto não ^{REQUIRED_SITUATION}] deve [ser aplicado ^{REQUIRED_SITUATION}] [por período superior a 3 semanas ^{TIME}], [especialmente em áreas cobertas por fraldas ^{PLACE}]. 'In infants and children up to 4 years of age, the product should not be applied for more than 3 weeks, especially in areas covered by diapers.'	Em lactentes e crianças abaixo de 4 anos, [o tratamento não ^{REQUIRED_SITUATION}] deve [prolongar-se ^{REQUIRED_SITUATION}] [por mais de 3 semanas ^{TIME}], [especialmente nas zonas cobertas por fraldas ^{PLACE}]. 'In infants and children under 4 years old, the treatment should not extend beyond 3 weeks, especially in areas covered by diapers.'
Tome os comprimidos por via oral com pouca quantidade de líquido não alcoólico. 'Take the tablets orally with a small amount of non-alcoholic liquid.'	[As gotas ^{REQUIRED_SITUATION}] devem [ser dissolvidas em um pouco de líquido não alcoólico ^{REQUIRED_SITUATION}]. 'The drops should be dissolved in a small amount of non-alcoholic liquid.'
[Você ^{REQUIRED_SITUATION}] deve [usar este medicamento exclusivamente nos olhos ^{REQUIRED_SITUATION}]. 'You should use this medication exclusively in the eyes.'	Este medicamento é de uso oftálmico. 'This medication is for ophthalmic use.'

Table 3 shows the case of the target *dever.v* in the sense of presenting a situation required for medication use. This was recurrent in the corpus, both in the texts addressing patients and those addressing healthcare professionals. Almost all automatic annotations assign *dever.v* to the *Desirable_event* frame, which suggests that a particular state of affairs is desirable.

Since these are medication package inserts, usage instructions for a medication are guidelines that must be followed and carefully observed by the patient. If they are not, negative consequences may occur. Given this context, the *Required_event* frame is the appropriate one, because in this frame the state of affairs obtained by the required situation prevents a negative consequence from occurring. This is the main difference between the *Desirable_event* and *Required_event* frames.

A further example is illustrated in Table 4.

Table 4: The polysemy of *diminuir.v* (*reduce.v*) / *aumentado.a* (*increased.a*)

PATIENT	HEALTHCARE PROFESSIONAL
Tome JANUMET durante as refeições, para diminuir [a possibilidade de [distúrbios estomacais _{ITEM} _{ATTRIBUTE}]]. 'Take JANUMET with meals to reduce the possibility of stomach upset.'	JANUMET deve ser administrado duas vezes ao dia, durante as refeições. 'JANUMET should be administered twice a day, with meals.'
Pacientes com insuficiência do fígado podem apresentar [risco de toxicidade _{ATTRIBUTE}] aumentado . 'Patients with liver failure may have an increased risk of toxicity.'	Pacientes com insuficiência hepática podem apresentar [risco de [toxicidade _{ITEM} _{ATTRIBUTE}]] aumentado , particularmente mielossupressão graus III-IV. 'Patients with hepatic insufficiency may have an increased risk of toxicity, particularly Grade III-IV myelosuppression.'

The frame assigned by the automatic parser in the sentences shown in Table 4 was *Cause_expansion*, in which an AGENT or a non-human CAUSE causes an ITEM to change its physical size. However, there is no physical size change in reducing the chance of stomach disorders. Therefore, the best-fit frame is *Change_position_on_a_scale*, as this frame indicates the change in position of an ITEM on a SCALE from an INITIAL_VALUE to a FINAL_VALUE.

5. Conclusions and Future Work

Our analysis of automatically annotated sentences revealed some of the challenges faced by an automatic frame parser. These range from tokenization and part-of-speech tagging to assignment of a correct frame. Some of the problems found, as is the case of tokenization and part-of-speech tagging, have an impact on downstream tasks, such as frame and frame element assignment. These cases are likely to be solved in the near future with better tokenization and part-of-speech models. Other problems, however, are more complex to solve, as they demand interpretation, as seen in cases of polysemy.

Overall, using automatic frame parsers seems promising and their output is illuminating regarding polysemous language and the need to further expand the frames available in FrameNet. We believe our work is relevant as a first look at the types of post-editing an automatic annotator requires. Human/manual annotation is a time-consuming process and qualitatively identifying the types of errors made allows for effective curation of automatic annotations in the near future of annotation.

As further steps in our research, we plan to conduct a new training of the parser based on the edits made and to perform post-editing on 200 new sentences to evaluate whether there were improvements, particularly in cases of polysemous lexical items.

Regarding the analysis of the language of medication package inserts, our results showed that the frames evoked in the texts addressing healthcare professionals and patients are similar, although the lexical structure is presented differently. Future work will be carried out to assess the level of semantic similarity between the two types of texts.

References

- Belcavello, F., Torrent, T. T., Matos, E. E., Pagano, A., Gamonal, M., Sigiliano, N., Dutra, L. V., Abreu, H. D., Samagaio, M., Carvalho, M., Campos, F., Azalim, G., Mazzei, B., Oliveira, M. F., Luz, A. C., Ruiz, L. P., Bellei, J., Pestana, A., Costa, J., Rabelo, I., Silva, A., Roza, R., Souza, M., & Oliveira, I. (2024). Frame2: A FrameNet-based Multimodal Dataset for Tackling Text-image Interactions in Video. *International Conference on Language Resources and Evaluation* (pp. 7429–7437). ELRA and ICCL.
- Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 8440–8451), Online. Association for Computational Linguistics.
- Das, D., Schneider, N., Chen, D., & Smith, N. A. (2010). Probabilistic Frame-Semantic Parsing. In *North American Chapter of the Association for Computational Linguistics* (pp. 948–956).
- Das, D., Chen, D., Martins, A. F., Schneider, N., & Smith, N. A. (2014). Frame-semantic parsing. *Computational linguistics*, 40(1), 9–56. doi:10.1162/COLI a 00163
- Dutra, L., Lorenzi, A., Larré, L., Belcavello, F., Matos, E., Pestana, A., Brown, K., Gonçalves, M., Herbst, V., Reinach, S., Teixeira, R., de Paula, P., Pellini, A., Sequeira, C., Sabino, E., Leal, F., Conde, M., Grespan, R., & Torrent, T. (2023). Building a Frame-Semantic Model of the Healthcare Domain: Towards the identification of gender-based violence in public health data. In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, (pp. 338–346). Porto Alegre: SBC. doi:10.5753/stil.2023.234145
- Fillmore, C. J. (1982). Frame semantics. In Linguistic society of Korea (Ed.), *Linguistics in the morning calm* (pp. 111–137). Hanshin Publishing Co.
- Fillmore, C. J. (1985) Frames and the semantics of understanding. In *Quaderni di Semantica*, 6(2), 222–254.
- Fillmore, C., & Baker, C. (2009). A frames approach to semantic analysis. In B. Heine, & H. Narrog (Eds.), *The Oxford Handbook of Linguistic Analysis* (pp. 791–816). Oxford University Press.
- Ruppenhofer, J., Ellsworth, M., Schwarzer-Petruck, M., Johnson, C. R., & Scheffczyk, J. (2016). *FrameNet II: Extended theory and practice*. International Computer Science Institute.
- Timponi Torrent, T., Matos, E., Belcavello, F., Viridiano, M., Gamonal, M., Costa, A. D., & Marim, M. C. (2022). Representing Context in FrameNet: A Multidimensional, Multimodal Approach. *Frontiers in Psychology*, 13, 1–20. <https://doi.org/10.3389/fpsyg.2022.838441>
- Viridiano, M., Lorenzi, A., Timponi Torrent, T., Matos, E. E., Pagano, A., Sathler Sigiliano, N., Gamonal, M., de Andrade Abreu, H., Vicente Dutra, L., Samagaio, M., Carvalho, M., Campos,

F., Azalim, G., Mazzei, B., Fonseca de Oliveira, M., Luz, A.C., Pádua Ruiz, L., Bellei, J., Pestana, A., Costa, J., Rabelo, I., Silva, A., Roza, R., Souza Mota, M., Oliveira, I., & Pelegrino de Freitas, M. (2024). Framed Multi30K: A Frame-Based Multimodal-Multilingual Dataset. *International Conference on Language Resources and Evaluation* (pp. 7438–7449). ELRA and ICCL.

Xia, P., Qin, G., Vashishtha, S., Chen, Y., Chen, T., May, C., ... & Van Durme, B. (2021). LOME: Large ontology multilingual extraction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations* (pp. 149–159), Online. Association for Computational Linguistics.

Zadeh, B. Q., Petruck, M. R., Stodden, R., Kallmeyer, L., & Candito, M. (2019). SemEval-2019 Task 2: Unsupervised Lexical Frame Induction. *International Workshop on Semantic Evaluation* (pp.16–30).

Acknowledgements

This work was supported by the National Council for Scientific and Technological Development (CNPQ) under grants 151361/2023-1, 315749/2021-0 and 313103/2021-6, and the Foundation for the Coordination and Improvement of Higher Education Personnel (CAPES) under grants 88887.936139/2024-00 and 8887.683333/2022-00.

We would like to thank the following junior researcher grantees from the State Funding Agency of Minas Gerais (FAPEMIG) and the National Council for Scientific and Technological Development (CNPq) who collaborated in compiling our corpus: Fernanda Teixeira Santana, Giovana Cristina Bastos de Oliveira, Jessica Victoria da Silva, João Pedro de Barros Faria Leite, Matheus Fernandes Marques and Miguel Novaes Marinho.

In addition, we would like to thank anonymous reviewers for their helpful comments.

Contact information

Maucha Gamonal

Federal University of Minas Gerais
mgamonal@ufmg.br

Adriana Pagano

Federal University of Minas Gerais, CNPq
apagano@ufmg.br

Tiago Torrent

Federal University of Juiz de Fora, CNPq
tiago.torrent@ufjf.br

Ely Matos

Federal University of Juiz de Fora
ely.matos@ufjf.br

Arthur Lorenzi

Federal University of Juiz de Fora
arthur.lorenzi@estudante.ufjf.br

Ellert Thor Johannsson and Thordis Ulfarsdottir

THE ROLE OF SEMANTIC FIELDS IN CONTEMPORARY ICELANDIC DICTIONARIES

Abstract This paper accounts for a system of semantic fields that was developed in Iceland around the turn of the century. The purpose of the system was to help describe the semantic properties of the Icelandic vocabulary and to be a practical tool in lexicographic work. The system categorizes words into semantic fields, enabling nuanced organization and practical applications in monolingual and bilingual dictionaries. The article details the system's origins, structure, and implementation, including its role in producing specialized glossaries and enhancing dictionary editing efficiency. While the system has proven valuable, some inconsistencies in classification and level of detail are noted, suggesting areas for improvement. Lastly closing observations are presented.

Keywords lexical semantics; semantic domains; semantic fields; Icelandic; dictionary editing

1. Introduction

The topic of this article is a specific system of semantic fields designed to describe lexical and semantic relationships within the vocabulary of Icelandic. When the Institute of Lexicography (later The Árni Magnússon Institute = ÁMI) in Reykjavík, Iceland, started compiling web-based dictionaries in 2004, this system was from the start integrated into the lexicographic work. This means actively using certain semantic information in dictionary editing and for other purposes such as producing specialized glossaries.

In this article, we account for this system, showcasing its properties and usefulness in lexicographic work. We give some examples of its practical use in contemporary lexicography where it has been employed in a monolingual Icelandic dictionary as well as several bilingual dictionaries. We divide our discussion into subsections. In Section 2, we give a background on Icelandic and online lexicographic and linguistic resources. In Section 3, we proceed to account for the system itself and how it is organized. We discuss key attributes of the system and its practical application. In Section 4, we demonstrate how the semantic fields function in practice in the lexicographic work and how they are used in dictionary editing. Furthermore, we show how the system can facilitate creation of specialized glossaries. We also mention some inconsistencies in the system that we are aware of. Finally, there are concluding remarks.

2. Background

Icelandic is a North-Germanic language with about 350 thousand native speakers. It is closely related to Norwegian, Danish and Swedish. Icelandic is the only official language in Iceland, and despite its relatively few speakers it is not considered endangered.

Icelandic is widely used in all aspects of daily life at all levels of society. Icelandic is the language of the parliament, the courts, the government and the municipalities, the schools at all levels and other institutions that engage in and provide public services. Icelandic authorities pursue an active language policy that aims to preserve Icelandic and ensure its continued prosperity in the digital age, e.g., promoting language standards and the coining of Icelandic neologisms (cf. Kristinsson, 2014; 2020).

The Árni Magnússon Institute serves a key role in implementing the language policy as a research center for the Icelandic language. The institute is entrusted with a range of language related roles and activities, such as preserving medieval Icelandic manuscripts and other valuable collections, as well as dictionary publishing and actively conducting scholarly activity related to the language and its text sources.

The lexicography of Icelandic has a rich tradition and is actively pursued today, mostly within the realms of ÁMI (cf. Úlfarsdóttir & Bjarnadóttir, 2017). In recent years the Institute's main lexicographic focus has been on modern web-based online dictionaries. Publications include ISLEX, a multilingual dictionary project involving Icelandic and five other Nordic languages, Danish, Swedish, Norwegian, Finnish and Faroese (cf. Úlfarsdóttir, 2014), *Íslensk nútímamálsorðabók* (Dictionary of Contemporary Icelandic (DCI)) (cf. Jónsdóttir & Úlfarsdóttir, 2019), which is the only monolingual dictionary of Icelandic that is continually updated, as well as several bilingual unidirectional dictionaries, e.g., Icelandic-French, Icelandic-English and Icelandic-Polish. Besides these online dictionaries the Institute manages other web-based linguistic resources that have been developed for Icelandic, such as a large database of terms within various disciplines (term bank), a database of inflectional patterns (BÍN), a reference website for orthographic guidelines, and more (cf. the web portal *malid.is* (Jónsdóttir et al., 2018)).

3. The Semantic System

3.1 Origin

At the core of the lexicographic endeavors of the past decades lies a system for semantic classification that was originally developed by Jón Hilmar Jónsson, an Icelandic linguist and lexicographer. This system arose from Jónsson's extensive work on collocations and fixed expressions where the meaning of every phrase was mapped out by a specific semantic label, or "concept," such as *anger*, *poverty*, *freedom*, *time span* and *distance*. Jónsson employed a system of semantic fields that is not connected to individual words, but to phrases and fixed expressions. The results of his efforts were published as printed onomasiological works (dictionaries of concepts) (Jónsson, 2002; 2005). These printed works later formed the foundation for the online resource *Orðanet* (*ordanet.arnastofnun.is*), which demonstrates the relationship between different concepts and the vocabulary that represents them in a wide variety of ways.

3.2 Semantic Fields

Jónsson's system of concepts was designed to categorize the lexis of the contemporary language and was developed from scratch. The process involved identifying and

classifying the semantic content of almost all lexical items (cf. Jónsson, 2002; 2005). His approach is rooted in the understanding of shared semantic features and conceptual relationships between the lexical items. His work provided a structured framework for organizing the Icelandic vocabulary, and in principle capturing both single-word entries and multi-word expressions. This system was designed only with Icelandic in mind and has no relation to other broadly used systems based on semantic criteria, like the well-known WordNet (cf. Feldbaum, 1998), which has been used to describe the semantic components of some related languages such as Danish (cf. Nimb et al., 2014).

The original system for semantic classification served as an impetus for continued lexicographic work. In the early 2000s, foundations were laid for compiling a new Icelandic–Nordic digital dictionary that later would become ISLEX (Úlfarsdóttir, 2014). As a first step in the process of starting this new project, a word list was compiled from existing sources, which was intended to serve as the basic lemmalist for the new dictionary. Jónsson took on the task of providing semantic labelling to this core vocabulary that was to serve as material for future dictionaries. Each word was assigned to one or more semantic fields and this information was registered with all the other lexical information as part of the dictionary database.

The resulting system employs a network of semantic fields. Each lemma is assigned to one or more of these fields depending on the number of senses that have different semantic properties. In the first version of the system there were over 6000 semantic fields defined, but many of those only contained 1 or 2 items (i.e., words) and were therefore of limited use in categorizing the lexis. About 3000 semantic fields can be considered useful in defining semantic information and of those around 1500 are significant as they contain ten or more items.

3.3 The System in Practice

To better account for the system, it is good to take a closer look at how it is employed in the lexical description of the Icelandic vocabulary in practice. For those purposes we will look at a particular lexicographic project, namely the monolingual dictionary DCI. This dictionary currently contains 56,000 words. All nouns, verbs and adjectives are categorized as belonging to one, two or more semantic fields. As a result, the vocabulary can be grouped in various ways and targeted in specific lexicographic work.

The data from DCI illustrates how the semantic information is associated with particular words and meanings. For example, the noun *gítar* ‘guitar’ is classified as belonging to only one field, namely “musical instrument” whereas *gulrót* ‘carrot’ belongs to two fields, “plant” and “food”. The word *brúðkaup* ‘wedding’ has a more nuanced meaning and is classified as belonging to several fields: “wedding”, “marriage”, “gathering” and “amusement”.

The semantic fields receive a heading (indicated here with double quotation marks “”). The heading usually consists of one or two descriptive nouns that represent the semantic characteristics of a particular field. As the headings are part of the regular vocabulary of

Icelandic, the words selected as headings also appear as headwords in dictionaries and as such are analysed for semantic properties like other words. As a result, the heading of a semantic field is often included as one of the words belonging to that particular field. For example, the semantic field “wedding” contains wedding-related words such as *bride*, *groom*, *wedding banquet*, etc. and also the event itself, i.e., *wedding*, which has been chosen as the representative heading for this particular semantic field.

There are great variations in how many words fall under each field, e.g., less than 10 words are part of the field “flag”, around 100 are found under “competition” and over 1000 are labelled “animal” or “medicine”. Table 1 gives an overview of the 25 largest semantic fields, i.e., those that have the most items.¹ For clarity, the names of the fields as well as the example words have been translated from Icelandic into English.

Table 1: The 25 largest semantic fields in the system

Field (Icelandic)	Translation	Words	Example words in translation
Læknisfræði	Medicine	1337	psychosis, optic nerve, screening
Matur	Food	1217	caviar, cardamom, chestnut
Mannlýsing	Personality trait	1110	naive person, bastard, drunkard
Dýr	Animal	1044	sheep, cat, goldfish, spider
Ástand	Condition	953	outdated, decrepit, adversity
Gras	Plant	939	oak, basil, dandelion
Starfsheiti	Profession	906	pharmacist, architect, chauffeur
Tími	Time/Period	837	advent, delivery time, teatime
Atvinnustarfsemi	Job related	818	farming, upholstery, bookseller
Tal	Speech/ talk	767	message, insinuation, reprimand
Fatnaður	Clothing	678	sock, uniform, close-fitting, rag
Húsbúnaður	Household item	600	armchair, toaster, spoon
Skóli	School	584	college, graduation, pupil
Lögfræði	Law	584	legacy, main proceedings, verdict
Eiginleiki	Characteristic	563	modesty, determination, aggressiveness
Fjármál	Finance	547	interest, current account, financial crisis
Tæki	Equipment/Tool	536	bread maker, winch, megaphone
Íþróttir	Sport	518	badminton, martial art, referee
Tónlist	Music	511	bass guitar, jazz, pop song
Líkamshluti	Body part	501	shoulder, nose, foot
Samfélag	Society	467	bourgeois, pensioner, private sector
Veður	Weather	429	headwind, storm, heatwave
Tölvur	IT/Computer	428	backup, bandwidth, update
Hljóð	Sound	424	footsteps, racket, squeak
Húshluti	Part of building	413	balcony, kitchen, main entrance

¹ The data from DCI refer to the state of the dictionary in 2022. As there are continuous additions and revisions made to the lemmalist, the numbers are subject to change.

As seen in Table 1, “medicine” and “food” are the semantic fields with the most numerous items. Closely following are “personality trait”, “condition”, “animal” and “plant”.

Nouns (40,000 lemmas) constitute about 70% of the vocabulary of DCI, so the numbers in Table 1 principally apply to nouns and their semantic properties. The other two big word classes are adjectives, 9,200 lemmas (16%), and verbs, 4,500 lemmas (8%). Regarding adjectives, their semantic fields tend to be different from those of the nouns. In many cases, they are more precise which often results in a number of small or very small groups, e.g., “loneliness” to which only three adjectives belong, and “recklessness” with six adjectives.

In addition, when Jónsson came up with the system, he first divided the adjectives into two main categories: “relating to persons” and “non-persons”. Each of these two main categories was then divided into smaller subcategories with a varying number of items. Examples of the semantic fields of adjectives relating to persons are “eye colour” (e.g., *blue-eyed*), “selfishness” (*selfish, egocentric*) and “financial status” (*wealthy, broke*). Non-person adjectives include “temperature” (*warm, cold*), “landscape” (*flat, mountainous*) and “usefulness” (*useful, useless*). Of course, many adjectives can either describe persons or inanimate things, in which case they were labelled as both, e.g., the semantic fields “messiness” (*messy, untidy*) and “strangeness” (*eccentric, odd*).

Obviously, each word class does not have its own specific semantic properties independent of other word classes. However, there are some definite trends, such as the semantic field “animal” which consists almost exclusively of nouns. Semantically, there is a great deal of overlap between lemmas belonging to different word classes, e.g., *deception* (noun), *deceptive* (adjective), *deceive* (verb); this also includes adverbs, *deceptively*. All these lemmas have been assigned to the semantic field “deception” as the heading for the field is always a noun.

Verbs have their own semantic characteristics, and their semantic fields tend to be some kind of action or activity, e.g., “change”, “movement” and “fighting”. In this context, “sleep” and “death” are also a form of activity. When the lemmas were processed (given a definition, etc.), it seemed natural to treat cognate words together as a single ‘package’: *sleep* (verb), *sleep* (noun) and *sleepy* (adjective). More words belonged to this semantic field, such as *slumber* and *doze*.

The sixty largest verbs (large as measured by their frequency in texts, the number of senses in the dictionary and the number of verbal phrases) were not processed in the same way as other lemmas. These verbs include (Icelandic verbs translated into their English equivalents) *take, come, go, say, stand* and *lie*. Instead of labeling the relevant lemmas with many different semantic field headings, these verbs were simply categorized as multisense lexical items. Due to their semantic complexity it turned out to be more effective in the lexicographic work to process them as a special group of lemmas separate from the main semantic field system, where the semantic properties of each verb could be analysed individually. Similar method was also applied to basic function words like prepositions, conjunctions and pronouns, where the general semantic classification simply did not work. Such words were processed based on formal properties of their particular part of speech.

4. Semantic Fields in Dictionary Editing

As is clear from the discussion above, the semantic field assignment creates manageable groups of words that share certain semantic traits. This has been extremely useful in practical dictionary work. In this section, we focus on the role and relevance of the system of semantic fields in the compilation of actual contemporary dictionaries that have been published in recent years or are currently being produced by ÁMI (cf. Section 2). All the above-mentioned dictionaries share a similar microstructure and a common source vocabulary that has been categorized by semantic fields.

4.1. The Semantic Fields as Part of the Lexical Data

The semantic field system was from the very start used to add important semantic information to the lexical data found in the dictionaries. Even if the classification originally focused mainly on phrases and other multi-word units, the semantic labelling was soon extended to span the whole vocabulary, as there was an ongoing and prolific development of bilingual dictionaries at ÁMI.

The following examples are from one of the unidirectional bilingual dictionaries, an Icelandic-English dictionary, which is currently being compiled and where this system is employed in the editing work. Figure 1 shows a simple lemma and demonstrates how various lexical information is conveyed.

<u>stíkilsber</u> n n	
1 FRAMB	ogg,mp3 framburður
2 BEYGING	→BEYGING
3 HLUTAR	STÍKILS-BER
4 SKÝRING	ber stíkilsberjarunnans
5 LATINA	(lat. Ribes uva-crispa)
6 MYND	
7 EN-jafn	gooseberry
8 MSVIÐ	ber/gras/ávöxtur/matur
9 STIG	I-1 E-1 MN

Fig. 1: The entry word *stíkilsber* ‘gooseberry’ in the Icelandic-English dictionary as it appears to the editor

When we look at the entry in Figure 1, we notice that apart from the lemma itself *stíkilsber* (at the top) it contains 9 types of data: 1) FRAMB pronunciation, given as a link to an audiofile; 2) BEYGING inflection, a link to the inflectional paradigm; 3) HLUTAR information on word division of compounds; 4) a definition in Icelandic: ‘ber stíkilsberjarunnans’; 5) Latin name of the plant; 6) illustration; 7) English equivalent, ‘gooseberry’; 8) four semantic field labels: “ber” ‘berry’, “gras” ‘plant’, “ávöxtur” ‘fruit’ and “matur” ‘food’; 9) STIG is an indicator of the stage of processing where E-1 means that the lemma has received an English translation and has been approved for next level of editing.

The semantic field label is placed near the bottom in each dictionary entry, as seen in Figure 1. In the dictionary writing system, the labels are defined as a special data type in the relational database, so that lemmas that share a label can easily be grouped together as a special data set which is relevant to the editorial work.

4.2 Editing Work Using Semantic Fields

4.2.1 Target Language Processing

By utilizing the semantic field system, editors can streamline the editing process, focusing more efficiently on semantically related words and concepts. The system allows editors to address interconnected lexical entries within shared semantic fields, thereby enhancing both accuracy and coherence in the definitions. This also ensures that an editor of a given target language does not repeatedly type in the same equivalent in the case of near synonyms. For example, Icelandic has several words meaning ‘endless, incessant’, like *óendanlegur*, *takmarkalaus*, *stöðugur* and *sífelldur*. By processing the target language by the semantic field “varanleiki” (‘durability, duration’), the editor of the English target language may choose to add a range of equivalent synonyms where they exist. In this case, he/she might use the words *endless*, *relentless*, *perpetual*, *ongoing*, *incessant*, *continual*, *constant*, *consistent*, *ceaseless*, *unremitting* and *uninterrupted*. This variety in the vocabulary is important because even though the dictionary is conceived of as a unidirectional dictionary, the interface allows the user to search for words not just belonging to the source language (Icelandic) but also the target language (English). This enhances the usefulness of the dictionary and opens it up to a broader user base.

4.2.2 Glossaries

The semantic field system can offer quite nuanced representation of lexical and semantic relationships. By identifying lexical and semantic correspondences between Icelandic and the paired languages, semantic details that exist in the vocabulary of non-Icelandic target languages are revealed. This means that glossaries can be generated focusing on specific semantic domains. So far, such efforts have resulted in the publication of bilingual glossaries of law (Icelandic-French and Icelandic-Spanish) and Economy and Finances (Icelandic-French). The Icelandic-French glossary of law contains 600 entries which all are part of the larger online dictionary LEXÍA. Glossaries of law are particularly important because the legal system has an urgent need for such aids, and Icelandic dictionaries usually do not have a great coverage of such terms. In court cases, Icelandic is always spoken, with the service of interpreters when needed, for example, when dealing with immigrants. Legal paperwork is also done in Icelandic which sometimes needs to be translated into other languages, normally by certified translators. These specific glossaries were compiled as online publications in pdf-format that could be downloaded for quick reference. They have also been used as teaching material in university courses in Iceland.

ákæra v
[fallstjórn: þolfall]
inculper
poursuivre en justice
<i>hún er ákærð fyrir fíkniefnasmygl</i>
elle est poursuivie pour trafic de stupéfiants
<i>hann hefur verið ákærður fyrir morð</i>
il a été inculpé de meurtre
ákærandi n m
procureur de la République
procureur général
ákærði adj
inculpé
<i>ákærði neitaði sök</i>
la personne inculpée a nié son crime

Fig. 2: Three entries from the Icelandic-French glossary of law (2018)

Besides this, some other specific data sets have been generated by using the semantic information. For example, a list of words related to wind and weather conditions was created at the request of the Icelandic Meteorological Office, with the intention of displaying the “weather word of the week” on their website.

4.3 Room for Improvement

As described in 4.2.1, the semantic fields have been of great help to the lexicographers during the editorial process of both the monolingual and the bilingual dictionaries. However, it is clear that the system does have some drawbacks that likely are due to its extremely short developmental phase and the fact that the system was initially implemented while still being worked on. This includes some structural issues we have identified which relate to inconsistencies in the semantic field assignment. It seems to us that many of these issues can be dealt with through simple adjustments and as a result the system as a whole would be improved. We will now discuss some examples of this.

The classification of words according to semantic fields has varying degree of nuances. In some respects, it is normal due to the lexical variety within the vocabulary, but in other cases it seems to be somewhat haphazard. This can be demonstrated by the category “animal”, which is one of the largest semantic fields. In many cases, both hypernyms (broader terms) and hyponyms (narrower terms) share the same semantic field label which then is only able to capture the broad semantic strokes. “Animal” may include individual species (*blackbird*) as well as their family (*thrush*) and the class (*bird*). In fact, the whole taxonomy of animals and plants is something that could be worked on to ensure better consistency.

In other instances, the classification is more fine-grained than necessary, so, for example, the noun *kokteill* ‘cocktail’ belongs to three semantic fields: “drinks”, “alcohol” and “blend”; and *hárgreiðsla* ‘hairstyle’ belongs to “hair”, “hairstyle”, “grooming” and “looks”. As is evident from the latter example, there are cases where hypernyms and hyponyms appear at the same level in the semantic field headings. Optimally the entire system could be revised to eliminate such apparent discrepancies.

Furthermore, in some cases we are aware of a certain lack of consistency in the scope of the semantic fields. Words denoting diseases (*plague, epilepsy, heart disease*) are assigned to the fields “disease”, “medicine” or a combination field “medicine/disease”, which appears superfluous. Words that refer to value generally have the semantic field label “verðmæti” (59 lemmas) but two lemmas are assigned to the field “dýrmæti”, which is synonymous to the main field and appears quite unnecessary. Such examples indicate that the names of the semantic fields themselves require a closer look and perhaps some further revisions.

5. Concluding Remarks

In this paper, we have accounted for the system of semantic fields developed for lexicographic work in Icelandic. Overall, the discussion has illustrated the significance of the semantic field system in organizing and enriching contemporary Icelandic dictionaries, particularly highlighting its role in enhancing the efficiency of dictionary editing processes and facilitating the identification of semantic correspondences between Icelandic and other languages. The discussion emphasizes the practical utility of the system for generating specialized glossaries, thereby contributing to a nuanced understanding of semantic relationships within and across languages. It is possible that the system could be useful in lexicographic work in other languages beside Icelandic, especially languages that share some structural similarities. However, it would require some revision and streamlining as the system in its current form has some known discrepancies and lacks uniformity when it comes to the level of details of each semantic field.

Language learners, researchers, and lexicographers benefit from the enhanced semantic classification that sheds light on various lexical relationships. This paper contributes to the ongoing discourse on lexicographical methodologies and highlights the relevance of language specific linguistic frameworks.

References

BÍN = *Beygingarlýsing íslensks nútímamáls*. Bjarnadóttir, K. (Ed.). The Árni Magnússon Institute for Icelandic Studies. Retrieved May, 2024, from bin.arnastofnun.is

DCI= Dictionary of Contemporary Icelandic. *Íslensk nútímamálsorðabók*. Jónsdóttir, H., & Úlfarsdóttir, Þ. (Eds.). The Árni Magnússon Institute for Icelandic Studies. Retrieved May, 2024, from islenkordabok.is

Fellbaum, Ch. (1998). *WordNet: An Electronic Lexical Database*. MIT Press.

Icelandic-English Dictionary/Íslensk-ensk orðabók. Úlfarsdóttir, Þ. (Ed.). The Árni Magnússon Institute for Icelandic Studies. Retrieved May, 2024, from enska.arnastofnun.is

ISLEX, Íslensk-norræn veforðabók [Icelandic-Nordic Dictionary]. Úlfarsdóttir, Þ. (Editor-in-chief). The Árni Magnússon Institute for Icelandic Studies. Retrieved May, 2024, from islex.is

Íslenskt-franskt fjármálaorðasafn [Icelandic-French glossary of finance]. (2020). Retrieved May, 2024, from <https://lexia.arnastofnun.is/media/uploads/fjarmalaord.pdf>

Íslenskt-franskt lögfræðiorðasafn [Icelandic-French glossary of law]. (2018). Retrieved May, 2024, from <https://lexia.arnastofnun.is/media/uploads/logfraedi2.pdf>

Íslenskt-spænskt lögfræðiorðasafn [Icelandic-Spanish glossary of law]. (2020). Retrieved May, 2024, from <https://lexia.arnastofnun.is/media/uploads/ISL-SPAENSKT-logfrordasafn.pdf>

Jónsdóttir, H., Kristinsson, A. P., & Steingrímsson, S. (2018). Málið.is: A Web Portal for Information on the Icelandic Language. In J. Čibej, V. Gorjanc, I. Kosem, & S. Krek (Eds.), *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts* (pp. 841–845). Znanstvena založba Filozofske fakultete Univerze v Ljubljani.

Jónsdóttir, H., & Úlfarsdóttir, Þ. (2019). Íslensk nútímamálsorðabók – kjarni tungumálsins. *Orð og tunga*, 21, 1–25. <https://doi.org/10.33112/ordogtunga.21.2>

Jónsson, J. H. (2002). *Orðaheimur*. JPV.

Jónsson, J. H. (2005). *Stóra orðabókin um íslenska málnotkun*. JPV.

Kristinsson, A. P. (2014). Ideologies in Iceland: The Protection of Language Forms. In A. K. Hultgren, F. Gregersen, & J. Thøgersen (Eds.), *English in Nordic Universities: Ideologies and Practices* (pp. 165–177). John Benjamins Publishing Company.

Kristinsson, A. P. (2020). Between Scylla and Charybdis: On Language Situation and Language Policy in Contemporary Iceland. *Bendriné kalba – Standard Language*, 93, 1–19. <https://journals.lki.lt/bendrinekalba/article/view/900/11>

LEXÍA. *Íslensk-frönsk orðabók* (Icelandic-French Dictionary). Úlfarsdóttir, Þ. (Editor-in-chief), Davíðsdóttir, R. E. (Editor of French). The Árni Magnússon Institute for Icelandic Studies. Retrieved May, 2024, from lexia.hi.is

Nimb, S., Trap-Jensen, L., & Lorentzen, H. (2014). *Den danske begrebsordbog*. Det Danske Sprog- og Litteraturselskab.

Úlfarsdóttir, Þ. (2014). ISLEX – a Multilingual Web Dictionary. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* (pp. 2820–2825). European Language Resources Association (ELRA).

Úlfarsdóttir, T., & Bjarnadóttir, K. (2017). The lexicography of Icelandic. In P. Hanks, & G.-M. de Schryver (Eds.), *International Handbook of Modern Lexis and Lexicography* (1–12). Springer. https://doi.org/10.1007/978-3-642-45369-4_104-1

WordNet. Retrieved May, 2024, from wordnet.princeton.edu

Contact information

Ellert Thor Johannsson

The Árni Magnússon Institute for Icelandic studies
ellert.thor.johannsson@arnastofnun.is

Thordis Ulfarsdottir

The Árni Magnússon Institute for Icelandic studies
thordis.ulfarsdottir@arnastofnun.is

Robert Krovetz

MORPHO-SEMANTICS AND DICTIONARY ENTRIES

Abstract This paper is about morphology and semantics, and about how their interaction is reflected in the choices made by lexicographers about dictionary entries. The paper discusses run-ons, zero-affix morphology, and the relationship between variant forms and lexical ambiguity. We compare run-on entries with variants that are headwords across different dictionaries. We found high agreement about the variants that were attested as headwords, and low agreement about the variants that were attested as run-ons. Corpus data showed differences between run-ons and headwords as well. We also compared a sample of variants that have an explicit affix with those that do not. We found many similarities with regard to the lexical semantic relationships that are involved. The paper concludes with a discussion of criteria for when derivational variants should be listed in a dictionary, and the opportunity presented by an electronic dictionary to teach a user about morphology and meaning.

Keywords morphology; lexical semantics; corpus linguistics

1. Introduction

By trade, he's a corn-chandler [said one character]. "And what on earth *is* a corn-chandler?" Peggy asked crossly...

Lady Mear said: A man who chandles corn, I suppose. Even my underrated intelligence can work that out.

--- Barbara Worsley-Gough (1932, p. 149)

Morphology is about the ways that words can vary in form. It includes a wide variety of phenomena: acronyms (*FBI*), abbreviations (*corp.*), initialisms (*C.B.S.*), slashonyms (*d/b/a - doing business as*), compounds (*blackbird*), inflectional variants (*dogs, loves/loved/loving*), derivational variants (*teacher, appropriation, misread*), back-formations (*bartend*), and zero-affix variants such as *cook* as a verb and as a noun. More than one type of variation can apply, e.g., *Federal Bureau of Investigation* can be referred to as *FBI* or *F.B.I.*

Morpho-semantics is about how different word form variations relate to meaning. According to Lieber (2012, p. 2108) "the most neglected area of morphological theory in the last three decades has been derivational semantics". Similarly, Levin & Rappaport Hovav (2017) observe that "The relationship between lexical semantics and morphology has not been the subject of much study".

Dictionaries are an excellent and understudied resource for investigating morpho-semantic relationships. They allow us to evaluate which derivational variants are

given explicit definition, and which have multiple meanings. We can also use corpus analysis to help identify highly-related variants. These variants often qualify as *run-ons*, which are derivational variants that are listed at the end of a homograph. We can compare such variants with those that are explicitly defined.

The notion of zero-affix morphology expresses the idea that a word can have an affix that is not visibly present. This can be seen with the word *undesirables*. Starting from the root *desire*, it transforms into *desirable*, and then *undesirable*. However, both *desirable* and *undesirable* are adjectives, and *undesirables* is a plural noun. There must be an implicit step that follows *undesirable* that converts the adjective to a noun. This is called *zero-affix morphology*, also known as *conversion*¹ or *functional shift*, because it involves a shift from one part of speech to another. Most of the literature focuses on the semantic relationships that are involved with zero-affix variants rather than their derivation. For example, *cook* as a noun is a person who cooks.

The semantic relationships between roots and derivational variants (as well as zero-affix variants) are very similar to the relationships seen with thematic roles. These relationships express different roles that are played by the noun phrases in a sentence: AGENT, INSTRUMENT, PATIENT, LOCATION, DESTINATION and others. For example, in the sentence *John ate the spaghetti with his fork*, the word *John* is an AGENT, and *fork* is an INSTRUMENT. If the sentence were *John ate the spaghetti with his wife*, John's wife plays the role of CO-AGENT rather than INSTRUMENT. These relationships are essential to understanding the meaning of the sentence. The suffix *-er* is often used to convey an AGENT or an INSTRUMENT relationship, as with *singer* or *shredder*. Similarly, *cook* as a noun is an AGENT (a person who cooks), and *ski* as a noun is an INSTRUMENT (something that is used for skiing). Morpho-semantics is linguistically interesting because it is neither entirely rule-governed nor completely idiosyncratic. We will discuss this further in Section 2.

Dictionaries differ in the types of morphological information they provide, and about how they convey this information to the user. The COBUILD Dictionary includes regular and predictable inflectional variants as part of the homograph for a word. Other dictionaries only include irregular inflectional variants (Macmillan), or define the irregular variant in a separate homograph (Collins). Dictionaries also vary about how they describe derivational variants. If the meaning is considered especially predictable, the derivational variant is listed as a *run-on* at the end of an entry. Sometimes this is expressed just as an affix (e.g. *-ly* at the end of the entry for *natural* in order to indicate that *naturally* is a word), and sometimes the variant form is spelled out.

This paper is an empirical analysis of the decisions that have been made by lexicographers about which word forms were included as run-ons. We conducted a corpus analysis using a large-scale resource of word families of derivational variants. We used the co-occurrences between roots and variants to help identify additional candidates for run-ons. We expect that those variants that co-occur will be more likely to be transparent.

¹ There are different viewpoints about conversion, and some make a distinction between conversion and zero-affix morphology (Valera, 2014).

We not only looked at run-on information, we also looked at whether derivational variants were expressed as run-ons or as headwords.

Finally, we looked at the relationship between derivational variants and word senses. Morphology is not a relationship between words, but between the senses of words. For example, *gravity* is related to *antigravity* and *microgravity*, but only in the sense of force-of-gravity. It is not related in the sense that means *serious*, as in *the gravity of the crime/offense*. We conducted an initial investigation about how often that occurs, and we make analogies between relationships that involve explicit affixes and relationships between zero-affix variants.

Our experiments were done with English. See Lieber (2012) for a discussion of derivational morphology in other languages.

The next section will discuss related work in linguistics and in lexicography. We will then describe the experiments that were done, and the results.

2. Related Work

The lexicon used to be considered “a repository of exceptions” (Bloomfield, 1933). This view changed with the publication of *Remarks on Nominalization* (Chomsky, 1970). That paper noted the similarities between verbs and nominalizations, such as *The enemy destroyed the city* compared with *The enemy’s destruction of the city*. The paper revolutionized the view of the lexicon by recognizing that it must have structure in order to support such regularities. It led to work on derivational morphology as a research subject, and its importance is reflected in a recent book entitled *Nominalization: 50 Years on from Chomsky’s Remarks* (Alexiadou & Borer, 2020).

Another seminal work is *The Categories and Types of Present-Day English Word Formation* (Marchand, 1969). That work is still a key reference for word formation and the interactions with semantics. In the second edition, Marchand makes analogies between zero-derivation and explicit affixes. So *clean(adj)/clean(v)* are related in the same way as *legal(adj)/legalize(v)*. In both cases the verb is an action that results in the state referred to by the adjective. Lipka says “if the concept of zero is really taken seriously, it is not the noun which denotes the agent, etc., but rather the zero-morpheme, since it is the very parallelism of *cheat/Ø* vs. *swindle/er*, *cook/Ø* vs. *bak/er*, *spy/Ø* vs. *observ/er* on which the theoretical concept of zero is based” (Lipka, 1975, p. 386). We want to automatically identify the type of relationship by extracting such information from dictionary definitions. For example, *wad* (verb) can be defined as “to make a wad of”. That is, the verb bears the semantic relationship of “formation” to the noun. This is similar to the effect that the morpheme *-ize* has on the noun *union* in order to make the verb *unionize*. Section 6 describes initial efforts to make these connections.

Princeton WordNet has been augmented with morphological information (Fellbaum & Miller, 2003). Princeton WordNet is the most widely used lexicon in Computational

Linguistics, and it consists of synsets² that express different types of lexical semantic relationships between word senses. Each synset is based on open-class parts-of-speech (nouns, verbs, adjectives, and adverbs). The morphological information adds connections across part-of-speech based on derivational relationships (e.g., *beauty/beautify*, *generalize/generalization*, *connect/connection*) and zero-affix relationships such as *cook*. Mitetelu et al. (2023) report on analyzing words that are zero-affix variants with regard to semantic classes that are at the top of the hierarchy in WordNet. These are classes like *noun.person*, *noun.artifact*, *verb.change*, *verb.motion*, and others. There are 25 noun classes and 15 verb classes in all. They found that zero-affix variants were highly frequent in their dataset, and it was the only affix that occurred with all 14 types of semantic relationships that the resource describes (Agent, Instrument, State, and others). They also found differences between zero-affix variants and explicit variants with regard to the distribution of the classes that were at the top of the hierarchy, and the semantic relationships between the words and variants.

Lieber (2004) mentions the following research issues in morpho-semantics:

- *The polysemy question.* This is the question about why an affix can express one type of relationship with one word, and a different type of relationship with another. For example, the way *-er* can sometimes be used to indicate an AGENT, and sometimes an INSTRUMENT.
- *The multiple-affix question.* This is the question about why different affixes are used to express similar relationships, as with *-ize* and *-ify* to indicate causative verbs, and *-er* and *-ant* to indicate agent nouns.
- *The zero-derivation question.* This is the question of how to account for word formation in the absence of an explicit affix.
- *The semantic-mismatch question.* This is the question of why there are morphemes that do not seem to add any meaning (e.g., *in* in *longitudinal*), and sometimes we have two affixes that seem to be redundant (e.g., *-ic* and *-al* in *dramatical*). Why doesn't adding an affix always add to the meaning? Why is it that *realistic* does not mean "pertaining to a realist"?

With regard to lexicography, DeCesaris (2021) mentions there has been relatively little research on morphological structure and dictionaries. She also notes that "the fine line between being entirely semantically transparent and only partially so is often blurred". What is added by an affix that warrants additional description?

The work of Sue Atkins with Charles Fillmore on FrameNet (Fillmore & Atkins, 2012), and the work of Patrick Hanks on Norms and Exploitations (Hanks, 2013) is especially relevant to the interaction of morphology and lexicography. FrameNet is a resource

² A synset is a group of sense-individuated words that share a common meaning. WordNet is a combination of a dictionary and a thesaurus.

that organizes the lexicon into a network of frames, where each frame represents a conceptual structure that captures the relationships among words and the roles they play in specific events or actions. FrameNet allows frames to inherit properties from more general frames, facilitating the capture of linguistic generalizations, while also permitting the expression of idiosyncratic information specific to particular lexical units. The work on Norms and Exploitations captures the idea that there are norms in the lexicon as well as creative ways that we can exploit those norms and therefore get irregularities.

The next section will discuss experiments to combine corpus analysis with dictionaries in order to get a better understanding of lexical semantics and the decisions made by lexicographers. This will be followed by a section about the results, and a section that compares how derivational variants are attested across three dictionaries. We then discuss our investigation of zero-affix variants. The paper will conclude with a discussion of the opportunities that are provided by electronic dictionaries for conveying morpho-semantic information.

3. Design of Experiments

We conducted an experiment with run-ons, and we made two initial investigations, one with derivational variants that were explicitly defined, and the other with zero-affix variants.

For the experiment with run-ons, we wanted to see if corpus information could be used to create such information automatically. We looked at co-occurrences within a paragraph to see how run-ons compared with headwords, and with variants that were neither run-ons nor headwords.

We used a download of the Wikipedia to compute co-occurrence information for two datasets. The first dataset used paragraphs, and the second used the entire article. We used a set of 5065 word families that we created for a different purpose for this work.³ We manually identified the run-ons for the homographs in the *Longman Dictionary of Contemporary English* (LDOCE) for all words that started with the letters A-C. We also identified those derivational variants that were explicitly defined. Our aim was to identify new candidate run-ons based on corpus data, and to see how run-ons and headwords compared.

For the investigation of zero-affix morphology, we used a second resource. The morpho-semantic links in Princeton WordNet include both derivational and zero-affix variants.⁴ We extracted the zero-affix variants from this dataset, and compared them with the first 100 derivational variants that are headwords in LDOCE. This work will be discussed in Section 6.

³ <http://lexicalresearch.com/resources/derivational-word-families.v12.tar>

⁴ See <https://wordnetcode.princeton.edu/standoff-files/morphosemantic-links.xls> for the resource file. The semantic relationships are discussed in (Fellbaum et al., 2007).

4. Run-ons and Headwords

There are 5065 word families in our resource, and 13,488 derivational variants. We used a subset of 1171 roots that began with the letters A-C for our comparison. There were 3147 variants for this set. Table 1 gives a breakdown of those variants with regard to whether they appear as a run-on, or as a suffixed headword, or as a prefixed headword in the LDOCE sample.

Our aim was to compare derivational variants that were run-ons with those that were explicitly defined, and to see how well we could identify variants that were presumed to be transparent. We used a subset of the roots that began with A-C because we manually identified whether the variant was attested as a headword or as a run-on. Prefixed variants were looked up under their first letter.

We found that of the 3147 variants, 601 were run-ons (19%), 989 were headwords (31%), and 1557 were candidates (50%). Of the headwords, 769 were suffixed variants, and 220 were prefixed variants. Of the candidates, 795 were suffixed and 762 were prefixed. Some of the variants did not occur in the Wikipedia download, some of the roots did not occur, and sometimes the variant and root did not co-occur in a paragraph. The rest of the run-ons, headwords, and run-on candidates were used to determine co-occurrences, and the number of variants for each group is shown in the Table (e.g., of the 601 run-ons, 457 co-occur with their roots within the context of a paragraph).

There were 439 candidates that did not co-occur with their roots within a paragraph. The co-occurrences can be low because morphological variants often occur with a lower frequency than the root, and in addition a paragraph is a relatively short context. We looked at two ways to capture more co-occurrences: 1) increasing the size of the context from a paragraph to an entire Wikipedia article, 2) using additional corpora. Table 1 compares the results using co-occurrence in a paragraph with co-occurrence in the entire article. We used two domain-specific datasets for supplementing the Wikipedia paragraph dataset: Medline, and Juris.⁵ In the Medline dataset the co-occurrences were within a title and abstract, and in Juris they were within a paragraph. We wanted to assess the impact of using datasets that were very different. The combined dataset increased coverage, but not as much as expanding the context from a paragraph to an entire Wikipedia article. Using the combined dataset there were 237 candidates that did not co-occur compared to 218 using the entire Wikipedia article and 439 using Wikipedia paragraphs.

⁵ The Medline dataset consists of over 17 million documents from the National Library of Medicine (<http://pubmed.ncbi.nlm.nih.gov>). The Juris dataset is made up of cases, statutes, briefs, treaties and other legal material (<https://public.resource.org/justice.gov/index.html>).

Table 1: A distribution of derivational variants into different classes. The variants are from a resource of word families. The classes are run-ons, suffixed headwords and prefixed headwords sampled from the *Longman Dictionary of Contemporary English* (LDOCE). The number of words and the percentages of the population that co-occur are given for the different classes, as well as how this varies for different datasets. The table also includes prefixed and suffixed variants that co-occur with the corresponding roots, but which are not run-ons or headwords (candidates for run-ons). The Don't Co-Occur row refers to candidates that do not co-occur with their root within the dataset

	Wiki Paragraph	Wiki Article	Medline	Juris
Run-on	457 (76%)	520 (86%)	371 (62%)	261 (43%)
Suffix HW	716 (93%)	745 (96%)	598 (78%)	512 (67%)
Prefix HW	209 (95%)	205 (93%)	168 (76%)	141 (65%)
Don't Co-Occur	439	218	418	500
Suffix Co-Occur	532 (70%)	632 (78%)	432 (54%)	194 (24%)
Prefix Co-Occur	494 (61%)	615 (76%)	487 (60%)	306 (40%)

We will discuss these results in Section 7. Table 2 shows the results over the entire word family resource, broken down by the type of affix. This was computed over paragraphs in the Wikipedia, but the results for the entire article are similar. We found that *-ly*, *-ion*, and *-er* were the most highly ranked suffixes based on productivity (number of derivational variants). The most productive prefixes were *un-*, *re-*, and *in-*. The Frequency column reflects the subset of the words in which the derivational form with that affix co-occurs with the corresponding root in the Wikipedia.

Table 2: The 10 most productive suffixes and prefixes for a set of over 5,000 word families. The table lists the number of words with each affix (how productive it is) for the roots and variants that co-occur within a paragraph in the Wikipedia dataset

Affix	Type	Frequency	Affix	Type	Frequency
ly	suffix	1295	un	prefix	825
ion	suffix	1049	re	prefix	370
er	suffix	983	in	prefix	217
ness	suffix	556	non	prefix	156
al	suffix	476	sub	prefix	142
ity	suffix	412	pre	prefix	108
ic	suffix	339	inter	prefix	106
able	suffix	318	dis	prefix	98
or	suffix	220	over	prefix	93
ist	suffix	205	de	prefix	82

We looked at the derivational variants and run-on entries in more detail. We wanted to know more about why lexicographers chose to explicitly define a variant. We also wanted to know about consistency. If the homograph had more than one sense, did the run-on apply to all of the senses? The first 100 homographs for derivational variants and the first 100 run-ons in LDOCE were assessed.

There were several cases where it was not clear why a derivational variant was explicitly defined. For example, consider these two definitions:

- (v) **acquiesce** - to agree, often unwillingly, without raising an argument; accept quietly
 (adj) **acquiescent** - ready to agree without argument

The homograph for the first word has a run-on of *-escence*, and the second has a run-on of *-ly*. It is not clear why the lexicographers could not have used *-escent*, *-escently* and *-escence* as run-ons to the first homograph. There were also cases where most of the senses correspond, but where one sense of the root does not have a corresponding sense with the variant, or vice-versa. Such differences could be mentioned in a usage note. Section 7 gives criteria for when derivational variants should be used as headwords.

We found that run-ons were generally consistent in that they applied to all of the senses in a homograph. However, there were exceptions. For example, the homograph for *acute* had run-ons of *-ly* and *-ness*. These apply to *acute* in the sense of illness. These run-ons do not apply to the usage *acute accent*, which was another sense that was listed in the same homograph.

5. Comparison With Other Dictionaries

In our experiment with run-ons, we expected that when a variant and a root co-occur, that would be an indication of semantic transparency. We expected that the population of run-ons would have more co-occurrences with their roots than headwords co-occurring with their roots. But, that is not what we found. Both suffixed and prefixed headwords co-occur more often. The candidates for run-ons also do not co-occur as often as the headwords. We examined two other dictionaries, *Collins COBUILD English Language Dictionary* (Sinclair, 1987) and *The American Heritage® Dictionary of the English Language* (Soukhanov, 1992) to see how they compared with the Longman dictionary (LDOCE) about how derivational variants were attested. We chose COBUILD because it is a British dictionary for learners of English (as is Longman). We chose the American Heritage® dictionary because it is an American dictionary with a focus on usage. We used three sets of 100 words each for the comparison: 1) the first 100 derivational variants in LDOCE that were headwords, 2) the first 100 derivational variants that were run-ons, 3) a set of 100 words from our word family resource that were neither headwords nor run-ons (i.e., words that were candidate run-ons). We wanted to know how these three sets were described in the other two dictionaries. All of these sets involved only suffixed variants.

We found high agreement between the dictionaries with regard to the first set. Of the 100 that were headwords, 89 were also headwords in COBUILD and 91 were headwords in the American Heritage® dictionary. For COBUILD, there were 3 that were attested as run-ons (*absently*, *acidity*, and *adoption*), and 8 that were not defined as headwords or run-ons (*abridgment*, *ablative*, *academician*, *accumulator*, *acidify*, *acidulated*, *acquaintanceship*, and *adoptive*). For American Heritage®, the other 9

variants were all run-ons (*absently, accountancy, acquaintanceship, acrobatic, actually, addictive, additional, administrative, and adoption*).

We found low agreement for the second set. Of the 100 run-ons in LDOCE, 21 were attested as headwords in COBUILD, and 13 were headwords in the American Heritage® dictionary. There were 29 that were also run-ons in COBUILD, and 80 that were run-ons in American Heritage®. There were 50 that were neither headwords nor run-ons in COBUILD, and 7 for the American Heritage® dictionary.

For the third set (candidates for run-ons), 9 were attested as headwords in COBUILD, and 28 as headwords in American Heritage®. There were 2 that were attested as a run-on in COBUILD, and 61 that were attested as run-ons in the American Heritage® dictionary.

There were significant differences in ambiguity between the three sets when they were attested as headwords in the other dictionaries. For the 89 members of the first set that were attested in COBUILD as a headword, 40 (45%) were ambiguous. For the 91 members that were attested in the American Heritage® dictionary as headwords, 71 (78%) were ambiguous. In contrast, for the 21 members of the second set (run-ons in LDOCE) that were attested as headwords in COBUILD, only 5 (24%) were ambiguous. Of the 13 that were attested in the American Heritage® dictionary as headwords, 8 (61%) were ambiguous. For the third set, of the 9 that were attested in COBUILD as headwords, none were ambiguous, and of the 28 that were attested in the American Heritage® dictionary, 11 (39%) were ambiguous. With each set, the percentage of ambiguous headwords decreases, and this occurs for both dictionaries.

6. Zero-Affix Morphology

We identified the words that were indicated as having a zero-affix relationship using the morpho-semantic links in WordNet 3.0. These relationships were compared against the relationships for explicitly defined derivational forms. We used the first 100 homographs for derivational variants from LDOCE for this purpose, which is the same set we used in the comparison with run-ons. The type of relationship is given in the resource.

We did indeed find a number of cases where the semantic relationship was the same. For example:

Agent:	advocate, affiliate, ally, alternate, associate, author abortionist, accompanist, accountant, actor, adaptor, adherent
Instrument:	airbrush, aquaplane, autoclave, ax/axe abrasive, accelerator, accumulator
Theme:	abstract, ad-lib, advance, affix, award acquisition

Location:	angle, address abrasion, accommodation
Process/Result:	account, ache, advantage, aggregate, alarm, arch, average abbreviation, abridgment, accumulation
Process/State:	alert, anger, awe acquiescent, addiction, admissible

In each of the above relationships, the first line shows instances where there is no explicit affix, and the second line has instances where the affix is explicitly given. There is an analogous relationship in each group to the corresponding words. For example, an *author* is someone who *authors*, just as an *adherent* is a person who *adheres*. We obtained the semantic relationship for zero-affix variants from the morpho-semantic resource. We used our own judgment with regard to the classification of the derivational variants, although sometimes they were also found in the resource.

Not all of the zero-affix variants were found in LDOCE with more than one part-of-speech. For example, *affiliate* was only defined as a verb, and *author* was only defined as a noun. In some cases our classification for the semantic relationship for the derivational variant differed from the one given in the resource. The word *addiction* was classified as an event rather than a state (the LDOCE definition was “the state of being addicted”). The resource used the category *Undergoer* for the words we list under *Theme*.

We are exploring automatic methods to create zero-affix variants from a dictionary and from a corpus. To create the resource from a dictionary we wrote a script to create links between the definitions of word senses. If there were two or more open-class words (lemmas) in common between the definitions of a word that differ in part-of-speech, we create a link. For example, *tonsure* is defined in WordNet 3.0 as:

- (n) **tonsure** (the shaved crown of a monk’s or priest’s head)
- (n) **tonsure** (shaving the crown of the head by priests or members of a monastic order)
- (v) **tonsure** (shave the head of a newly inducted monk)

The verb, *tonsure*, has the lemmas *shave*, *head*, and *monk* in common with the first sense of the noun, and *shave* and *head* in common with the second sense. This overlap creates a link between *tonsure* as a noun and as a verb, and that link is used to create a dataset of zero-affix candidates. For identifying candidates from a corpus, we tagged a corpus (Wikipedia), and identified cases in which a word co-occurs in a paragraph with two different parts-of-speech. The more often it co-occurs, the more likely the word meanings are to be related despite the difference. For example, we expect that *cook* as a verb and as a noun co-occur more often than *train* as a verb and as a noun.

Zero-affix morphology is a research topic. Our aim in these comparisons is to use dictionaries to explore the semantic relationships involved. We believe that corpus data will complement what we can extract from dictionaries.

7. Discussion

The interaction between morphology and semantics presents challenges as well as opportunities for lexicography. When do we need to provide a derivational variant as a headword? There are at least the following cases:

1. When there is no apparent root. These are cases like *altruism*, *masochism*, and *ventriloquism*. Such words are also strongly connected with *-ist* variants, which have a transparent and predictable relationship to the root. We also have *ambiguous*, with the related form *ambiguity*.
2. When there is a high semantic distance between the root and the variant. For example, *social* and *socialism*.
3. When the variant has a more specific meaning than the root and one of the arguments of the root is “hardwired”. An example is the word *abstain*, which has variants *abstention* (abstain from voting), and *abstemious* (abstain from food and drink). The root form does not specify what the person is abstaining from, but this is made explicit with the variants.⁶
4. When the derived form is ambiguous. For example, *gravity* can be used in the sense *force-of-gravity*, or to mean *serious* (as in *the gravity of the crime*). This involves ambiguity between a regular word and a derivational variant. Another example is *invalid* referring to a person or to mean *not valid*. A second type of ambiguity is that between adjectival participles and tensed verbs. For example, *accomplished* can be used in the context *an accomplished pianist* or in the context *he accomplished the task*. The tensed verb use is expected, but the adjectival participle sense is not. It is not always clear how to divide a word’s usages into senses, and these are some of the cases that were encountered.

We note that it is also possible to have ambiguity between a root and an inflected form, as with *bats/nuts/crackers/bananas* as an adjective that means *crazy* or as a plural noun, or with *minutes* to refer to a singular noun (the *minutes of the meeting*) or a plural noun. The unpredictable usage needs to be explicitly defined.

We were able to use our resource of word families to suggest additional run-ons. Variants that were not attested as headwords or as run-ons in LDOCE were attested as run-ons in two other dictionaries. But we found low agreement between dictionaries about which variants should be described as run-ons. Sometimes they were attested as headwords, and sometimes they were not attested at all. In contrast, there was high agreement that the variants that were attested as headwords in LDOCE were attested as headwords in the other two dictionaries.

We also found a difference between run-ons and headwords in the experiment with corpus data. A greater proportion of the headwords co-occur with their roots

⁶ We consider it analogous to the process of Currying, in which an argument to a function is hardwired (as in converting *add(x,y)* into *add_one(x)*, which adds one to x rather than y). *Abstention* can also be used in the more general sense of a voluntary decision not to act.

compared with run-on entries co-occurring with their roots. This was true across four different datasets. We looked at the corpus data in more detail to determine why this was so. The suffixed variants that were attested as headwords were, on average, four times more frequent in the Wikipedia than the variants that were attested as run-ons. The suffixed run-on candidates were even less frequent. This pattern held true across the domain-specific corpora, although the degree of variation differed by corpus. The pattern was also true for prefixed forms. The run-ons for prefixed headwords were much less frequent, on average, than the headwords they corresponded with.

There were differences in ambiguity between headwords, run-ons, and run-on candidates. The percentage of ambiguous variants is greatest in the variants that are attested as headwords, and it is the least in the run-on candidates for LDOCE that are attested as headwords in the other dictionaries. This makes sense given the findings with frequency—more frequent words also tend to be more ambiguous. We need an additional way to determine if a derived form is ambiguous. As future work, we intend to look at Transformers (Vaswani et al., 2017) for this purpose. Transformers are a type of neural-network architecture in which the representation used for a given word form is context-sensitive. We also plan to use this approach to re-assess the relatedness of variants and roots for run-ons and run-on candidates.

The availability of dictionaries in electronic form presents opportunities for teaching the user about morpho-semantics. DeCesaris (2021) noted that the Random House dictionary presented combining forms⁷ as a list of words in which the combining form occurs. We can extend this to word families. When a user asks for a word to be defined, we can show the other members of the word family as well. Table 3 shows some of the largest word families in our resource. The word family for *state* is a combination of variants that are related to *state* as a verb and to *state* as a noun, and since they have different meanings they are arranged in different groups. There are different meanings for the nouns as well. The aim is to give the user an appreciation for the wealth of variation rather than just expose them to a run-on with a predictable meaning.

We should not only show the word family that the variant belongs to, we should also contrast instances from similar affixes when appropriate. For example, *non-* and *un-* both express negation, but they cannot be used interchangeably, and there are important differences in meaning: *nonchristian* refers to a person, but *unchristian* refers to a belief or a behavior. If a user asks about one, the other can be used as a contrast. The prefixes *de-* and *dis-* can also convey negation, but they are often associated with reversal of an existing state (as with *deactivate* or *disconnect*). In addition, *de-* can be used to indicate removal (e.g., *debone*), or reduction (e.g., *devalue*), among other meanings. Prefixed variants are generally not as frequent as suffixed variants, so English learners might not encounter them often enough to fully grasp the distinctions.

The contexts that are used to illustrate a word's usage can be chosen to include roots that co-occur with variants (as with our experiments). This can help inform the user about how different word forms occur in context.

⁷ A *combining form* is a morpheme that comes from Latin or Greek; they can often be paraphrased with an open-class word (e.g., *geo-* (Earth), or *hydro-* (water)).

Table 3: A sample of some of the largest word families in English

interpret	interpretation misinterpret misinterpretation reinterpret reinterpretation interpreter interpretable interpretability uninterpreted uninterpretable interpretive interpretively
polar	polarity multipolarity polarize polarizer polarizable polarizability polarization depolarize depolarization repolarize repolarization unpolarized nonpolar
state	stateless statelessness stateful stative statehood tristate multistate substate superstate interstate intrastate statement overstate overstatement understate understatement misstate misstatement restate restatement unstated

8. Conclusion

Dictionaries vary a great deal with regard to the morphological information they contain, and how it is conveyed. But we found a great deal of commonality with regard to the derivational variants that were defined as headwords. About 90% of a sample of variants from the Longman dictionary were also headwords in two other dictionaries. In contrast, we found low agreement with regard to variants that were run-on entries, which are provided without definition at the end of a homograph. They were sometimes explicitly defined in other dictionaries, sometimes they were run-ons, and sometimes they were not attested as either a headword or as a run-on.

We analyzed corpora for variants attested as headwords to see how they compared with variants attested as run-ons. We found that a greater proportion of headwords co-occurred with their roots compared with run-ons co-occurring with their roots. Headwords were much more frequent than run-ons, and those were more frequent than run-on candidates. We also found that there were differences in ambiguity when headwords, run-ons, or run-on candidates were attested as headwords in other dictionaries. Variants that were attested as headwords had the highest percentage of words that were ambiguous, and variants that were run-on candidates had the lowest percentage of ambiguous words.

While it is a common practice to include run-on information in dictionary entries, it is unclear that this is useful to the user. A comparison with a resource of word-families indicated many variants that could have been included as run-ons, but were not, and a sample of the variants that *were* defined included a number of cases where the meaning seems transparent. We provided a set of criteria regarding when a variant should be explicitly defined. We propose that it is beneficial to expose a user to entire word families rather than just run-ons, and that dictionaries should be set up to convey distinctions between variants with semantically similar affixes.

References

- Alexiadou, A., & Borer, H. (Eds.). (2020). *Nominalization: 50 Years on from Chomsky's Remarks*. Oxford University Press.
- Bloomfield, L. (1933). *Language*. Henry Holt.

Chomsky, N. (1970). Remarks on Nominalization. In R. Jacobs, & P. Rosenbaum (Eds.), *Readings in English Transformational Grammar* (pp. 184–221). Ginn & Co.

Choudhari, D., Damani, O., & Laxman, S. (2011). Lexical Co-occurrence, Statistical Significance, and Word Association. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* (pp. 1058–1068). Association for Computational Linguistics.

DeCesaris, J. (2021). Dictionaries and Morphology. In Z. Gavrillidou, L. Mitits, & S. Kiosses, (Eds.), *Lexicography for Inclusion: Proceedings of the 19th Euralex International Conference* (pp. 577–584). Democritus University of Thrace.

Fellbaum, C., & Miller, G. A. (2003). Morphosemantic links in WordNet. *Traitement automatique de langue*, 44(2), 69–80.

Fellbaum, C., Osherson, A., & Clark, P. E. (2007). Putting Semantics into WordNet's "Morphosemantic" Links. In *Proceedings of the Third Language and Technology Conference*, Poznan, Poland.

Fillmore, C., & Atkins, B. T. (2012). Towards a frame-based lexicon: the semantics of RISK and its neighbors. *Frames, fields, and contrasts* (pp. 75–102). Routledge.

Hanks, P. (Ed.). (1979). *The Collins English Dictionary*. Collins.

Hanks, P. (2013). *Lexical Analysis: Norms and Exploitations*. MIT Press.

Levin, B., & Rappaport Hovav, M. (2017). Morphology and Lexical Semantics. In A. Spencer, & A. Zwicky (Eds.), *The Handbook of Morphology* (pp. 248–271). Wiley.

Lipka, L. (1975). Review of: An Introduction to Modern English Word-Formation. *Lingua*, 37(4), 382–389.

Lieber, R. (2004). *Morphology and Lexical Semantics*. Cambridge University Press.

Lieber, R. (2012). Semantics of Derivational Morphology. In C. Maienborn, K. von Heusinger, & P. Portner (Eds.), *Semantics: An International Handbook of Natural Language meaning*, Vol. 3, (pp. 2098–2119).

Marchand, H. (1969). *The Categories and Types of Present-Day English Word Formation: A Synchronic-Diachronic Approach*, second edition. Beck.

Mititelu, V. B., Leseva S., & Stoyanova, I. (2023). Semantic analysis of verb–noun zero derivation in Princeton WordNet. *Zeitschrift für Sprachwissenschaft*, 42(1), 181–207.

Procter, P. (Ed.). (1978). *The Longman Dictionary of Contemporary English*. Longman.

Rundell, M. (2002). *Macmillan English Dictionary for Advanced Learners*. Macmillan.

Schone, P., & Jurafsky D. (2001). Is knowledge-free induction of multiword unit dictionary headwords a solved problem?. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

- Sinclair, J. (Ed.). (1987). *Collins COBUILD English Language Dictionary*. Collins.
- Soukhanov, A. (Ed.). (1992). *The American Heritage Dictionary of the English Language*, Third Edition. Houghton-Mifflin.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30(1), 261–272.
- Valera, S. (2014). Conversion. In R. Lieber, & P. Štekauer (Eds.), *The Oxford Handbook of Derivational Morphology* (pp. 154–168). Oxford University Press.
- Worsley-Gough, B. (1932). *Public Affaires*. Victor Gollancz.

Contact information

Robert Krovetz
Lexical Research
rkrovetz@lexicalrsearch.com

Haniva Yunita Leo

DO INDONESIAN SPEAKERS FEEL ‘PAIN’? NSM and Corpus-Based Approach to the Cross- Linguistic Concept of ‘Pain’ in Bahasa Indonesia

Abstract This study challenges the global notion of ‘pain’ by exploring the semantic nuances of pain-related concepts in Bahasa Indonesia and comparing them to English. It examines four lexemes: *sakit*, *nyeri*, *pedih*, and *perih* using the Natural Semantic Metalanguage (NSM) and a corpus-based approach. The study proposes alternative definitions to address the circularity and vagueness in the Great Indonesian Dictionary. Analyses reveal distinctions in physical and emotional conceptualizations: physical sensations differ in body locus, intensity, and duration, while emotional pain-like concepts relate to social relationships and vary in triggers, duration, and reactions. This highlights significant cross-linguistic distinctions between Bahasa Indonesia and English in conceptualizing ‘pain’.

Keywords pain; cross-linguistic; Bahasa Indonesia; English; NSM; corpus-based approach

1. Introduction

The perception and expression of pain vary significantly across cultures. Cross-cultural studies reveal that the concept of ‘pain’ is not universal (Goddard & Ye, 2016; Wierzbicka, 2012; 2016). In Bahasa Indonesia, words like *sakit*, *nyeri*, *pedih*, and *perih* convey different aspects of ‘pain’. Although medical literature often uses *nyeri* to mean ‘pain’ based on the International Association for the Study of Pain (IASP) definition, non-medical discourse from the BPPT-PAN English-Indonesian corpus sourced from CQP Web (Hardie, 2012) shows overlap in the usage of *sakit* and *nyeri*.

1. If someone suffers an acute **chest pain** associated with a heart attack, they call the ambulance

Jika seseorang menderita nyeri dada akut yang terkait serangan jantung, segera panggil ambulans.

2. ...play again in Jan. 2007 but continued to feel **pain** in his knee and went under the knife for a...

...untuk bermain lagi pada Januari 2007 tetapi terus merasa sakit pada lututnya dan kembali menjalani pembedahan ...

Both *sakit* and *nyeri* are commonly used to describe pain, but the Great Indonesian Dictionary (KBBI)(Badan Bahasa, n.d.) does not clearly distinguish between the words and other related concepts. The KBBI defines pain-related terms using a traditional *genus + specifica* approach, which often leads to vagueness and circularity. To illustrate, here are the definitions of each lexeme along with their English translations.

sakit a berasa tidak nyaman di tubuh atau bagian tubuh karena menderita sesuatu (demam, sakit perut, dan sebagainya)

‘feeling uncomfortable in the body or part of the body due to suffering from something (fever, stomach ache, etc)’

nyeri 1 a berasa sakit (seperti ditusuk-tusuk jarum atau dijepit pada bagian tubuh); rasa yang menimbulkan penderitaan

‘feeling painful (like being pricked by a needle or being pinched in a part of the body); feeling that causes suffering’

pedih a berasa sakit, seperti luka dicuci dengan sublimat

‘feeling painful as if the wound was washed by sublimate’

perih a pedih

‘smart/sting/sore’

The KBBI defines *sakit* with an open-ended and vague description, using terms like *etc.*, that obscure its specific causes. This vagueness extends to using *sakit* as a broad category that fails to clearly distinguish between various pain-related concepts, such as *pedih*, which might not be considered as suffering from something. Additionally, the definitions of pain-related words like *nyeri*, *pedih*, and *perih* are circular and complicated to understand. Users need to comprehend *sakit* to understand these other terms, but since *sakit* itself is vaguely defined and circularly explained through examples like *sakit perut*, it remains unclear and confusing.

To address these definitional challenges, the present study aims to explore the semantic nuances of the four related concepts. Following this, the natural semantic metalanguage (NSM) (Wierzbicka, 2012; Goddard & Wierzbicka, 2014) and corpus-based approach are introduced (Section 2) to deal with definitional challenges. Then, the pain-related concepts are analysed based on physical and emotional sensations along with the proposed explications (Section 3) before presenting the conclusion and further steps from the present study (Section 4).

2. Methods: NSM and Corpus-Based Approach

This study employs the Natural Semantic Metalanguage (NSM) for meaning analysis, a lexical semantics approach initiated by Wierzbicka in 1972. NSM posits that all languages share a core mini-language of lexical and grammatical elements, enabling neutral cross-cultural comparisons (Ye, 2019, p. 133). It uses reductive paraphrase to decompose word meanings into basic elements called semantic primes, with their manifestations in any language termed exponents. This study uses English exponents and their translations in Bahasa Indonesia.

The data for this study are sourced from two accessible Indonesian corpora: IndonesianWAC (Kilgarriff et al., 2010), obtained through Sketch Engine, and LCC Indonesian 2023, made accessible via CQPweb (Hardie, 2012). The profiles of these two corpora are presented in Table 1.

Table 1: Profile of Corpora

	LCC Indonesian 2023	IndonesianWAC
Year	2008--2022	2013
Size	573,557,097-word tokens	100 million words
Texts	mixed texts (mixed genres) collected from the Internet from 2012–2014, web-based news, and Wikipedia	texts collected from the Internet
Access via	https://cqpweb.lancs.ac.uk/lccindonesianv3/	https://app.sketchengine.eu

Analysis on semantic nuances is conducted based on significant collocates determined by high frequency in two corpora. In the Indonesian WAC, collocates are measured using Log Dice and extracted using Word Sketch and Word Sketch Difference, with outcomes depicted in graphs. In the LCC Indonesian 2023, collocates are measured using Dice Coefficient, which combines significance and strength of connection. These collocates, excluding unrelated instances like *rumah sakit* ‘hospital’, are visualized in graphs using the *ggplot* package in R (R Core Team, 2022).

Additionally, colloquial context analysis is derived from patient-doctor interactions on the Alodokter website (Kementerian Kesehatan Republik Indonesia, n.d.), filtering topics with relevant keywords. This results in 18 topics for analysis. While corpora and Q&A sessions are used to analyse the physical aspect of ‘pain’, the emotional sphere is examined through concordances and Indonesian song lyrics from the 1990s to 2000s, with 12 lyrics selected based on relevant keywords.

3. Results

This subsequent section presents the results of analysis of the four lexemes. The discussion is divided into physical and emotional aspect of ‘pain’-related concepts in Bahasa Indonesia.

3.1 Physical Pain-Related Concepts

To delve into the meaning of the four lexemes, the collocates of the lexemes from the left and right context (presented in Figures 1 to 4) are analysed and their contexts of use are re-examined from their concordances. The result of analysis is discussed based on locus, intensity, and duration of pain.

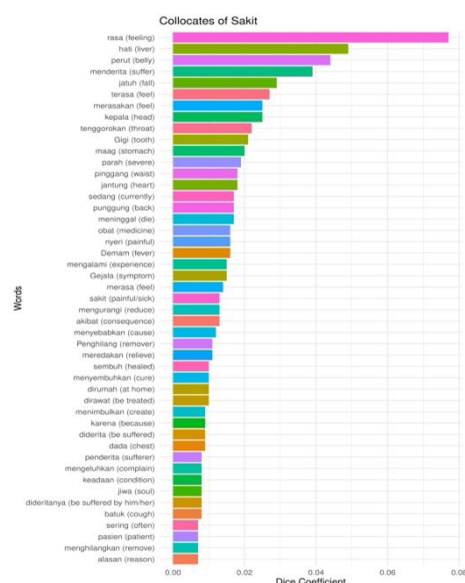


Fig. 1: Collocates of Sakit from LCC Indonesian 2023

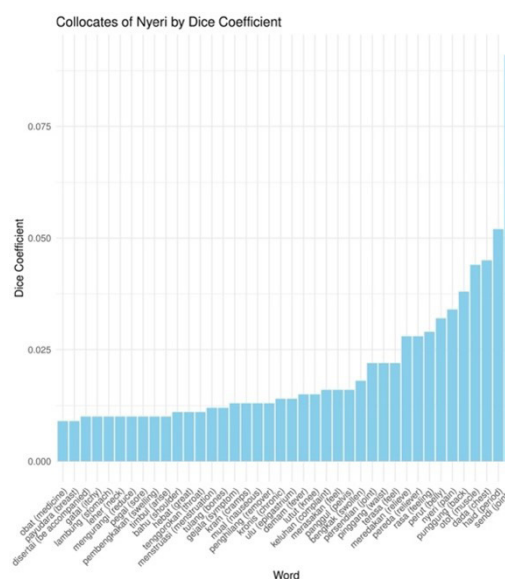


Fig. 2: Collocates of Nyeri from LCC Indonesian 2023

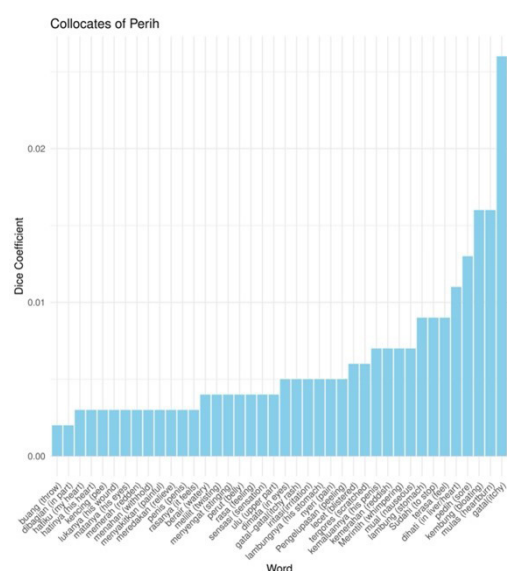


Fig. 3: Collocates of Perih from LCC Indonesian 2023

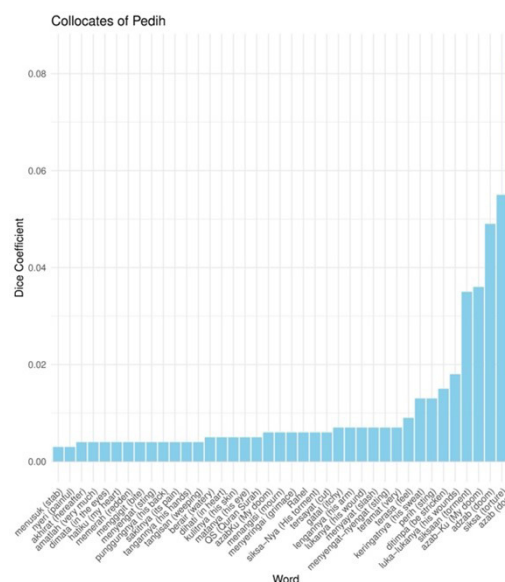


Fig. 4: Collocates of Pedih from LCC Indonesian 2023

In Figures 1 to 4, 40 high frequency collocates of the four lexemes from the left and right contexts are depicted. For example, *sakit* is frequently collocated with *rasa* 'sensation/feeling', *jatuh* 'fall', and body parts such as *hati* 'liver' and *kepala* 'head'. *Nyeri* co-occurs predominantly with body parts such as *sendi* 'joint', *dada* 'chest', and *otot* 'muscles'. *Perih* and *pedih* are collocated with terms such as *gatal* 'itchy', *luka* 'wound', and *mata* 'eyes'. Additionally, *azab* 'doom' and *siksa* 'torment' only co-occur with *pedih* in Islamic texts, which are excluded from the analysis.

When comparing Figure 1 and Figure 2, notable distinctions in terms of locus among the four related concepts emerge from their significant collocates. While *sakit* (Figure 1) frequently co-occurs with *rasa* 'feeling; sensation', *menderita* 'suffer', *jatuh* 'fall', and *terasa* 'feel', *nyeri* predominantly pairs with body parts such as *sendi* 'joints', *dada* 'chest',

and *otot* 'muscles', indicating its localization as best illustrated in (3). In this example, when describing *nyeri*, a specific body locus is necessary. However, when referring to the general sensation of 'pain' felt in the body, the term *sakit* in its derivative form *kesakitan* 'be in pain' is used since there is no similar derivative form for *nyeri* such as *?kenyerian*.

3. ...*demam* kemudian terjadi ***nyeri otot*** terutama
 ...fever then happen **pain muscle** especially
 di daerah kaki dan ketika
 in area leg and when
 penderita merasa ***kesakitan***.
 sufferer feel **be in pain.**
 '...fever, then **muscle pain** occurs, especially in the leg area and when it is held the patient feels **pain**.' (LCC Indonesian 2023)

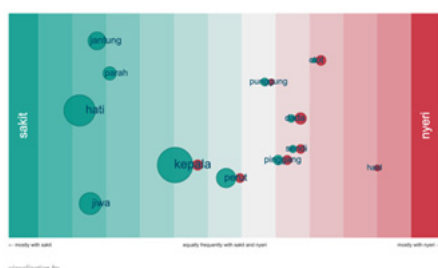


Fig. 5: Comparison of collocates between *sakit* and *nyeri*

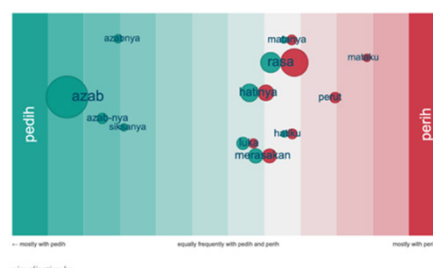


Fig. 6: Comparison of collocates between *pedih* and *perih*

As *nyeri* is typically localized compared to *sakit*, the collocates involving specific body parts in the right context (Figure 2) exhibit the specific locus and the nature of both lexemes. In terms of locus, the figure depicts that *nyeri* is mostly followed by terms like *sendi* 'joints' and *otot* 'muscles' which are less commonly occurred with *sakit* that is frequently followed by *kepala* 'head' and *perut* 'belly' (see Figure 5). The specific locus suggests internal locus in part of body compared to a more general sensation in *sakit* as expressed by patient's explanation quoted from Alodokter (4).

4. *Halo dok, telinga saya bengkak dan bagian dalamnya terasa nyeri. Rasanya mengganggu sekali dok, sekarang rasanya seperti tersumbat. obatnya apa ya dok? rasanya nggak nyaman bgt soalnya. Thanks*

'Hello Doc, my ear is swollen, and it **hurts inside**. It's really annoying, now it feels like it's clogged. What's the cure, Doc? It doesn't feel very comfortable. Thanks.' (Alodokter)

In contrast, *pedih* and *perih* are externally localised depicting by their significant collocates such as *gatal* 'itchy' and *luka* 'wound' (Figures 3 & 4) that represent sensation on skin. While both lexemes share external locus, these two words differ in the context of internal loci, as the collocates *lambung* 'stomach' and *perut* 'belly' only co-occur with *perih* (Figure 6).

5. Penyakit Maag merupakan penyakit akibat jaringan lunak lambung yang disebabkan adanya luka atau peradangan di lapisan lambung yang menyebabkan mulas atau **perih** di bagian perut akibat produksi asam lambung yang berlebihan.

‘Ulcer disease is a disease caused by the soft tissue of the stomach which is caused by injury or inflammation in the stomach lining which causes heartburn or **pain** in the stomach due to excessive production of stomach acid.’ (LCC Indonesian 2023)

6. Ambil busa sabun biarkan menempel pada kulit sambil digosok pelan-pelan maksimal 1 menit, lalu bilas hingga bersih, jika **kulit** terasa **pedih** atau gatal kurangi dosis pemakaian.

‘Take the soap foam, let it stick to the skin when rubbing it gently for a maximum of 1 minute, then rinse thoroughly, if the **skin** feels **sore** or itchy, reduce the dosage.’ (LCC Indonesian 2023)

As demonstrated in the previous example (6), it is crucial to have prior contact with specific body locations to experience *pedih*. This is in stark contrast to *perih*, which is primarily influenced by the disease and does not necessitate such contact (5).

In terms of intensity and duration, all pain-like sensations indicate unpleasant feelings to some degree, which the individual does not want, except for *pedih* due to its short-term nature (as in 6). When considering intensity, example (8) demonstrates *nyeri* as a physical sensation that the body cannot tolerate, unlike *perih* and *pedih*, therefore requiring *obat pereda nyeri* meaning ‘pain reliever’. Meanwhile, *sakit*, as illustrated in (7), is a severe sensation that wholly affects the individual.

7. Ia mengambil cuti setelah menderita **sakit**
 S/he took leave after suffer **ache**
perut yang berhubungan dengan stres.
 stomach which is related with stress.
 ‘She took a leave after suffering from **stomach ache** that was related to stress’. (IndonesianWAC)

8. Seseorang yang menderita **nyeri sendi** akan merasakan peradangan yang sangat hebat **serta rasa nyeri yang sangat tajam** yang tidak bisa dihindari.
 ‘Someone who suffers from **joint pain** will feel very intense inflammation and a **very sharp pain** that cannot be avoided.’ (LCC Indonesian 2023)

In terms of duration, *sakit* and *perih* take some time to recover due to their whole-body effect and disease-affected factor. On the other hand, *nyeri* is experienced in terms of frequency due to its unavoidable nature, such as *nyeri haid* or menstrual pain (Figure 5). Unlike other pain-like concepts, *pedih* is a short-term experience and is not typically seen as a problem that needs to be solved. This is evident from the less frequently discussed topic with a doctor on the Alodokter website.

The prototypical concepts of the four lexemes are constructed in the following explications with their Indonesian counterparts. The explications present three components as the prototypical scenario of 'pain': "feeling something bad in part of the body (a), thinking that something bad is happening to that part (b), and wanting it not to be like this (d)" (Goddard & Wierzbicka, 2014, p. 132).

[A1] <i>Sakit</i> in <i>Dia sakit perut/kepala/gigi</i> (She/he has stomach/head/tooth-ache) (a) someone feels something bad in his body (b) like someone can feel when it is like this: (c) something bad is happening to a part of this someone's body for some time (d) because of this, this someone feels something bad in his part of body for some time (e) this someone thinks like this: "I don't want to feel something like this, I want to do something because of this I cannot think like this: I can do something with my body"	[A2] <i>Dia sakit perut/kepala/gigi</i> (a) seseorang merasakan sesuatu yang buruk di badannya (b) seperti seseorang bisa rasakan waktu seperti ini: (c) sesuatu yang buruk sedang terjadi pada satu bagian dari badan orang ini untuk sementara (d) karena ini, orang ini merasakan sesuatu yang buruk di bagian badannya untuk sementara (e) seseorang ini berpikir seperti ini: "Aku tidak mau merasakan sesuatu seperti ini, aku mau buat sesuatu karena ini Aku tidak bisa berpikir seperti ini: Aku bisa buat sesuatu dengan badanku"
[B1] <i>Nyeri</i> in <i>sendinya terasa nyeri</i> 'his joints feel painful' (a) he feels something very bad in a part of his body (b) like someone can feel when it is like this: (c) something very bad is happening inside one part of this someone's body (d) because of this, this someone's feels something very bad inside the part of his body, not one time (e) this someone thinks like this: 'I know I can't think: I can't feel something like this'	[B2] <i>Nyeri</i> in <i>sendinya terasa nyeri</i> (a) dia merasakan sesuatu yang sangat buruk pada salah satu bagian badannya (b) seperti seseorang bisa rasakan waktu seperti ini: (c) sesuatu yang sangat buruk sedang terjadi di dalam salah satu bagian dari badan seseorang ini (d) karena ini, seseorang ini merasakan sesuatu yang sangat buruk di dalam bagian badannya, bukan satu kali (e) seseorang ini berpikir seperti ini: 'Aku tahu aku tidak bisa berpikir: Aku tidak bisa merasakan sesuatu seperti ini'
[C1] <i>Pedih</i> as in <i>matanya terasa pedih</i> 'his eyes felt sore' (a) someone felt something bad in one part of his body [eyes, skin] (b) like someone can feel when it is like this: (c) something touched his part of body (eyes) a short time before (d) because of this, something bad happened in this part of the body (e) at the same time, he felt something bad in his part of the body for a short time	[C2] <i>Pedih</i> as in <i>matanya terasa pedih</i> (a) seseorang merasakan sesuatu yang buruk pada satu bagian badannya [mata, kulit] (b) seperti seseorang bisa rasakan waktu seperti ini: (c) sesuatu menyentuh bagian badannya sesaat sebelumnya (d) karena ini, sesuatu yang buruk terjadi pada bagian badannya (e) pada saat yang sama, dia merasakan sesuatu yang buruk pada bagian badannya untuk sesaat
[D1] <i>Perih</i> as in <i>Perutnya (matanya; kulitnya) terasa perih</i> 'Her stomach (eyes; skin) feels sore' (a) she feels something bad in one part of her body [stomach, eyes, skin] (b) like someone can feel when it is like this: (c) this someone feels something bad in one part of her body for some time (d) because something bad has happened to this part of her body (e) when this someone feels something like this, she thinks like this: "I don't want to feel something like this"	[D2] <i>Perih</i> seperti <i>Perutnya (matanya; kulitnya) terasa perih</i> (a) Dia merasakan sesuatu yang buruk pada satu bagian badannya [perut, mata, kulit] (b) seperti seseorang bisa rasakan waktu seperti ini: (c) seseorang ini merasakan sesuatu yang buruk pada satu bagian badannya untuk sementara (d) karena sesuatu yang buruk sudah terjadi pada bagian badannya (e) ketika seseorang ini merasakan sesuatu seperti ini, dia berpikir seperti ini: "Aku tidak mau merasakan sesuatu seperti ini"

As a point of comparison, Wierzbicka's (2012) proposed explication is presented below.

- a) she felt something bad at that time
- b) like someone can feel when it is like this:
- c) something bad is happening in a part of this someone's body
- d) this someone feels something bad in this part of the body because of this
- e) this someone can't think like this at this time: "I don't want this"

Comparing the English concept of pain with the four related concepts in Bahasa Indonesia reveals three notable distinctions. First, the unitary component differs. Pain in English is unitary (as shown in component (a) in [E]), while in Bahasa Indonesia, the concepts are primarily physical because the emotional sensation takes a different locus *hati* 'liver' as the centre of emotion. Second, components (c) in explications [A1] and [D1] and (d) in [B1] and [C1] imply a temporal process, a feature absent in [E]. Lastly, although component (e) in English and Bahasa Indonesia's explications share the pain as an 'unwanted sensation', they are different in terms of the experiencer's control over the body and unavoidable sensation as presented respectively in components [A1] and [B1].

Unlike the unitary feeling in English, in Bahasa Indonesia, the emotional sensations take *hati* ‘liver’ as the locus, forming fixed expressions: *sakit hati*, *pedih hati*, and *perih hati*. This is evident from the significant collocates in all pain-like concepts, except for *nyeri*, which is mainly physical.

Due to limited occurrence in corpus data, only collocates of *sakit hati* are presented below.



The following examples illustrate the notable distinctions among the three emotional pain-like concepts in terms of triggers, reactions, and duration.

9. ***Sakit hatiku tiada terobati***

*Walau sejuta senyum kau berikan
Demi langit dan bumi aku bersumpah
Dirimu tak mungkin lagi kumaafkan
'My sakit hati is incurable
Even though you give me a million smiles
By heaven and earth I swear
I cannot forgive you'
(Meggie Z, Sakit Hati)*

10. ***Hatinya*** ***pedih*** *tersayat-sayat, tak tahan*
liver.his painful shredded not stand
menyaksikan *isak* *tangis dan duka-lara istrinya.*
see sob cry and sorrow wife.his
'His heart ached and tore apart; he couldn't bear to witness his wife's sob
and sorrow.'
(LCC Indonesian 2023)

11. ***Pedih hati saat kehilangan darah daging sendiri dan benar kata emak, itulah takdir untukku tapi emak hadir memayungiku dan inilah hikmah***
It hurts when I lost my child and it's true what my mother said, that's my fate, but my mother is here to console me and **this is the lesson...**
(LCC Indonesian 2023)

12. *Begitu berat beban hidupku
Ku menangis sendiri
Dia yang ku cinta telah pergi
Tuhan ku tak sanggup lagi
Menahan perihnya hati
Mengapa aku begini
'My life's burden is so heavy
I cried alone
The one I love is gone
God, I can't take it anymore
To Endure (my) heartache
Why am I like this'
(Perih, Ariel Noah)*

While all emotional pain is triggered by close social relationships such as lovers or spouses, *sakit hati* is mainly caused by shattered expectations that end up with *dendam* 'resentment' (Figure 7 & example 9). Differently, experiencing *pedih hati* and *perih hati* requires the component of loss and being abandoned (10–12). While both emotional sensations share these components, only *pedih hati* includes thinking about something bad that has happened or will happen to someone the experiencer cares for while feeling powerless to do anything about it (10).

When comparing the reactions, it is clear that *sakit hati* prompts a reaction from the experiencer toward others, while in *pedih hati* and *perih hati*, the focus remains in the experiencer's internal state.

Although the emotional sensations differ in trigger and reaction, they share the component of intensity but are different in duration. Unlike the varying degrees of physical 'pain,' all emotional sensations constitute a very distressing state cantered in *hati* 'liver.' However, they are different in terms of duration. While time may fade away the feeling of loss as the experiencer gains some insight from the sad event (11), resentment remains incurable (9).

All the components of emotional pain-like concepts are deconstructed into the following explications.

[F1] <i>Sakit hati</i> as in <i>seseorang merasa sakit hati</i> 'someone feels sakit hati' (a) someone feels something very bad for a long time (b) like someone can feel when he thinks about someone else like this: (c) "I know this someone will not do bad things to me This someone did something very bad to me" (d) when someone thinks something like this, he feels something very bad towards someone else (e) because of this, this someone wants to do something very bad towards someone else	[F2] <i>Sakit hati</i> seperti <i>seseorang merasa sakit hati</i> (a) seseorang merasakan sesuatu yang sangat buruk untuk waktu yang lama (b) seperti seseorang bisa rasakan waktu dia berpikir tentang seseorang yang lain seperti ini: (c) "Aku tahu seseorang ini tidak akan melakukan hal-hal yang buruk kepadaku Seseorang ini melakukan sesuatu yang sangat buruk kepadaku" (d) waktu seseorang berpikir sesuatu seperti ini, dia merasakan sesuatu yang sangat buruk terhadap seseorang yang lain (e) karena ini, seseorang ini mau melakukan sesuatu yang sangat buruk terhadap seseorang yang lain
[G1] <i>Pedih hati</i> as in <i>Pedih hatinya melihat anaknya terbaring lemah tak berdaya</i> 'it hurts her heart to see her child lying weak and helpless' (a) she feels something very bad for some time at this time (b) like someone can feel when she thinks about someone else like this: (c) "I feel something very good when I think about this someone, (d) I want this someone to be with me for a long time (e) I want very good things to happen to this someone" (f) At this time she thinks like this: (g) "Now I know something bad has happened to this someone (h) maybe this someone cannot be with me for a long time (i) I don't want this to happen to this someone (j) I want to do something now, I know I can't do anything now" (k) because of this, she feels something very bad for some time	[G2] <i>Pedih hati</i> seperti <i>Pedih hatinya melihat anaknya terbaring lemah tak berdaya</i> (a) Dia merasakan sesuatu yang sangat buruk untuk sementara pada saat ini (b) seperti seseorang bisa rasakan waktu dia memikirkan seseorang yang lain seperti ini: (c) 'Aku merasakan sesuatu yang sangat baik waktu aku memikirkan seseorang ini, (d) Aku mau seseorang ini ada denganku untuk waktu yang lama (e) Aku mau hal-hal yang sangat baik terjadi pada seseorang ini" (f) Pada saat ini dia berpikir seperti ini: (g) "Sekarang aku tahu sesuatu yang buruk sudah terjadi pada seseorang ini (h) Mungkin seseorang ini tidak bisa ada denganku untuk waktu yang lama (i) Aku tidak mau ini terjadi pada seseorang ini (j) Aku mau melakukan sesuatu sekarang, aku tahu aku tidak bisa melakukan apa pun sekarang" (k) karena ini, dia merasakan sesuatu yang sangat buruk untuk sementara

[H1] <i>Perih hati</i> as in <i>perih hatinya karena kehilangan istrinya</i> 'his heart aches because he lost his wife' (a) he feels something very bad at this time (b) like someone can feel when he thinks about someone else like this: (c) "I feel something very good when this someone is with me (d) I thought before, this someone could be with me for a long time (e) now, I know, this someone cannot be with me for a long time" (f) because of this, he feels something very bad for some time	[H2] <i>Perih hati</i> as in <i>perih hatinya karena kehilangan istrinya</i> (a) dia merasakan sesuatu yang sangat buruk pada waktu ini (b) seperti seseorang bisa rasakan waktu dia berpikir tentang seseorang yang lain seperti ini: (c) "Aku merasakan sesuatu yang sangat baik waktu seseorang ini ada denganku (d) Aku pikir sebelumnya, seseorang ini bisa ada denganku untuk waktu yang lama (e) sekarang, Aku tahu, seseorang ini tidak bisa ada denganku untuk waktu yang lama" (f) karena ini, dia merasakan sesuatu yang sangat buruk untuk sementara

The explications consist of some components. Component (a) highlights the difference between physical and emotional sensation. The absence of bodily locus in (a) is typical of emotional 'pain'. Component (b) presents the social relationship as a trigger following its prototypical scenario in indented parts of the explications. Components (d), (e), (f), and (k) present the reaction and duration.

4. Conclusion and Further Steps

This paper unpacks the semantics of 'pain'-related concepts in Bahasa Indonesia and compares the concepts with that of 'pain' in English. By using NSM and corpus-based approach, the present study proposes the explications of pain-related concepts in Bahasa Indonesia as alternative definitions to tackle the vagueness and circularity from the existing dictionary definitions. The study also highlights the difference in physical and emotion sensations, discriminating this concept with the unitary concept of pain in English.

The current study, while contributing to cross-linguistic research on the concept of 'pain' and providing potential solutions for creating definitions of culture-specific words, has two limitations. First, the proposed explications may be too complex for laypeople to understand due to their length and formulation. Therefore, if NSM is used to develop dictionary entries, concise formulations should be considered. Second, the results of this study are based on generated data from written text, and thus do not account for individual differences in experiencing and expressing 'pain' in both languages.

Despite the limitations, NSM explication can be applied to the domain of emotion and other related concepts whose meaning cannot be fully expelled using a *genus+specifica* type of definition. Developing dictionary entries of culture-specific words using NSM explication in a more concise formulation may benefit the L2 Indonesian learners in terms of understanding culture through words.

References

- Badan Bahasa. (n.d.). *Kamus Besar Bahasa Indonesia*. Retrieved May 20, 2023, from <https://kbbi.kemdikbud.go.id>
- Goddard, C., & Wierzbicka, A. (2014). *Words and Meanings Lexical Semantics across Domains, Languages, and Cultures* (First Edition ed.). Oxford University Press.
- Goddard, C., & Ye, Z. (2016). Exploring “happiness” and “pain” across languages and cultures In C. Goddard, & Z. Ye (Eds.), *“Happiness” and “Pain” across Languages and Cultures* (pp. 1–18). John Benjamins Publishing Company.
- Hardie, A. (2012). CQPweb—combining power, flexibility and usability in a corpus analysis tool. *International journal of corpus linguistics*, 17(3), 380–409.
- Kementerian Kesehatan Republik Indonesia. (n.d.). *Alodokter*. Retrieved September 26, 2023, from <https://www.alodokter.com/komunitas/diskusi/penyakit>
- Kilgariff, A., Reddy, S., Pomikálek, J., & Avinesh, P. (2010). A Corpus Factory for Many Languages. In N. Calzolari et al. (Eds.), *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA).
- R Core Team. (2022). R: A language and environment for statistical computing In R Foundation for Statistical Computing (Ed.). Vienna, Austria.
- Wierzbicka, A. (2012). Is Pain a Human Universal? A Cross-Linguistic and Cross-Cultural Perspective on Pain. *Emotion Review*, 4(3), 307–317. <https://doi.org/10.1177/1754073912439761>
- Wierzbicka, A. (2016). “Pain” and “suffering” in cross-linguistic perspective. In C. Goddard, & Z. Ye (Eds.), *“Happiness” and “Pain” across Languages and Cultures* John Benjamins Publishing Company.
- Ye, Z. (2019). The semantics of emotion: From theory to empirical analysis. In S. Pritzker, J. Fenigsen, & J. Wilce (Eds.), *The Routledge Handbook of Language and Emotion* (pp. 132–153). Routledge.

Acknowledgements

As this paper is part of my thesis for the Master of General and Applied Linguistics (Advanced) degree from the Australian National University, I would like to thank Dr Zhengdao Ye for her supervision. I also thank the institutional support from the Indonesian Endowment Fund for Education Agency and Language Development and Cultivation Agency, The Ministry of Education, Culture, Research, and Technology, Indonesia.

Contact information

Haniva Yunita Leo

East Nusa Tenggara Language Office, Language Development and Cultivation Agency, The Ministry of Education, Culture, Research, and Technology, Indonesia
hanivayunita@gmail.com

Michael Nguyen and Peter Juel Henriksen

THE LOCUM HYPHEN

A Formal Approach to the Lexicalization of Multiword Expressions with Rich Internal Semantics

Abstract We present a study of Danish multiword constructions containing one or more hyphens, such as *gas- og vandmester* ('gas- and water.repairman'; 'plumber'), *ilt- og brintatomer* ('oxygen- and hydrogen atoms') and *haveborde og -stole* ('garden tables and -chairs'). Although materially analogous, such constructions exhibit different semantics, falling — as we shall argue — into two distinct groups ("locum" vs. "pseudo locum" hyphen constructions). This study employs the COR₁ database (Jervelund et al., 2012; Widmann, 2024) and draws on the CLINK formal framework for computational lexicography (Henriksen, 2023; 2024a). The aim of this paper is to demonstrate how linguistic analysis of such multiword expressions with rich internal semantics can benefit from methods of computational lexicography — and vice versa.

Keywords hyphen; multiword expressions; language technology; structural ambiguity; punctuation; COR₁

1. Introduction

This paper presents a lexicographical study on Danish (written) multiword expressions (MWEs) containing hyphens preceded or followed by a blank space. Examples are *gas- og vandmester* ('gas- and water.repairman', 'repairman of gas and water'; 'plumber'), *ilt- og brintatomer* ('oxygen- and hydrogen atoms') and *haveborde og -stole* ('garden tables and -chairs'). We introduce the term *locum phrase* for a subclass of such phrases, viz. those that can be transformed into a synonymous phrase by replacing its hyphen(s) with lexical material occurring in the phrase. The hyphens of a locum phrase are termed *locum hyphens*. As we shall see, not all MWEs with hyphens are locum hyphens.

As lexical base, we employ the newly published Centrale Ordregister (Central Word Register alias "COR₁", cf., Henriksen, 2024a), a machine-readable projection of Retskrivningsordbogen, the official Danish Orthographic Dictionary (Jervelund et al., 2012). Retskrivningsordbogen (RO) defines the Danish orthographic norm and is published by The Danish Language Council (Act 1997). The Danish public sector (including the school system) is obliged to follow the norm. RO consists of a traditional dictionary (approximately 64,000 lemmas with morphological information) and a manual of 62 articles with the rules of Danish punctuation (comma marking, above all), abbreviation, capitalization, compounding and more.

COR₁ is specially aimed at the Danish NLP (natural language processing) sectors, serving as the lexical foundation of applications for spell and grammar checking, speech technology, machine translation, chatbots, AI and so forth.

As in most European languages, the hyphen grapheme (RO § 57) has several functions in Danish orthography. In this paper, we focus on the use of the hyphen in phrases where it replaces a part of a word (see § 57.2), for instance:

- (1) vin- og ølflasker
 wine- and beer.bottles
 ‘wine bottles and beer bottles’

As mentioned, we refer to such phrases as *locum phrases*. Lexicalizing MWEs such as locum phrases is challenging, not only in COR₁ but in computational lexicography in general.

In this paper, we discuss the semantic interpretation of the locum phrase as opposed to the *pseudo locum phrase* – a phrase that is materially identical to the locum phrase. Our analysis points to some possible improvements of the rule formulations in RO regarding both types of phrases.

2. RO and COR₁

Each word form in COR₁ carries a unique ID in this format:

COR. <lemma index> . <inflexion index> . <variant index>

COR₁ covers all RO lemmas (including a number of MWEs) and also contains sub-verbal lexemes (e.g., affixes, symbols and digits) and non-verbal lexemes (e.g., punctuation marks). Thus, all text elements with an individual semantic contribution (words and punctuation marks alike) carry their own COR-IDs:

Table 1: Samples from the machine-readable COR₁ dictionary

Lexeme	COR-ID	Gloss
bord	COR.44636.120.01	‘table’
borde	COR.44636.122.01	‘tables’
og	COR.00099.970.01	‘and’
-	COR.XP.017	(‘hyphen’, as used in this paper)

As for the rules of RO, they are worded rather informally in order to be user-friendly; thus, intuitive definitions are preferred over formal specifications. Often, large numbers of examples are provided from which the reader can infer how the rules are to be applied. In Section 3, we discuss in greater detail the locum hyphen and also introduce the pseudo locum hyphen.

3. The Locum Hyphen

In Danish, compounds and derivatives are typically written as one text word without any blank spaces or hyphens, e.g., *havebord* (‘garden table’) and *brugervenlig* (‘user-

friendly’). According to the rules of RO, a hyphen may substitute a part of a word (§57.2) (authors’ translation):

“A hyphen is used to show that two or more compounds or derivatives have a common part which is only mentioned once.”¹

This common part of the word is recoverable from the immediate syntactic context, more specifically from a conjunctive phrase (the locum phrase). Example (2) below is from RO (§57.2).

- (2) haveborde og -stole
 garden.tables and -chairs
 ‘garden tables and garden chairs’

The hyphen substitutes *have* (‘garden’), thereby qualifying as a locum hyphen. The phrase can be expanded without a change of semantics: *haveborde og havestole* (‘garden tables and garden chairs’). Note furthermore that both the substituted part (*have*, ‘garden’) and the non-substituted part (*borde*, ‘tables’) are lexical morphs².

3.1 Complex Cases of the Locum Hyphen

However, a number of complications arise in other cases. See (3)-(5), all likewise from RO, §57.2:

- (3) over- eller underskud
 over- or under.shot
 ‘surplus or deficit’

- (4) im- og eksport
 im- and ex.port
 ‘import and export’

- (5) A- og B-skat
 A- and B-tax
 ‘A-tax and B-tax’

In (3), the hyphen substitutes *skud*, which has no semantics in and of itself, and so it is no morph, and thus has no lexical entry.

In (4), both the substituted *port* and the non-substituted *im* are likewise not morphs, and thus not represented in any lexical entry.

¹ “Bindestreg bruges for at vise at to eller flere sammensatte eller afledte ord har en fælles del som kun bliver nævnt én gang.”

² We follow Henrichsen (2024a) in using the term *morph* for any text element that has an individual semantic contribution. The term includes but is not limited to morphemes, punctuation marks and linking elements.

In (5), what is substituted is not a string of letters but a morph + another hyphen (i.e., *-skat*). This is revealed by the expansion of the phrase: *A--skat og B-skat*. Double hyphens are, however, not used in Danish orthography.

Another complicated case is given in (6) (not mentioned in §57.2):

- (6) *køb.s.pris* *og* *-dato*
 buy.LINK.price *and* *-date*
 ‘(the) price of buying and (the) date of buying’

The hyphen substitutes *købs*, but *købs* has no lexical entry. This is because the *s* is not part of the morph *køb* ‘buy’ but a so-called linking element, which appears in many compounds in Danish (see e.g., Bauer, 1975, pp. 222–245).

3.2 Introducing the Pseudo Locum Hyphen

Although the examples (3)–(5) from §57.2 are complex cases for a parser, they (as well as example (6)) are in principle consistent with the translated rule passage above.

See, however, the MWEs in (7) and (8), taken from the lexical entries of RO:

- (7) *gas- og* *vandmester*
 gas- and *water.repairman*
 ‘repairman of gas and water’; ‘plumber’
- (8) *social- og* *sundhedsassistent*
 social- and *health.LINK.assistant*
 ‘social and health care assistant’

Expanding (7) shows that the hyphen does not substitute any part of a word: ??*Gasmester og vandmester* (‘repairman of gas and repairman of water’). Although the expanded phrase is structurally well-formed, there is no such thing as a *gasmester* (‘repairman of gas’) or a *vandmester* (‘repairman of water’).

Furthermore, it can be observed that the two conjuncts can neither be switched around (??*vand- og gasmester*, ‘water- and gas.repairman’), nor can the conjunction be replaced with another conjunction (??*gas- eller vandmester*, ‘gas- or water.repairman’). The same restrictions apply to (8).

Examples like (7) and (8) should be distinguished from superficially similar examples like (9) (a constructed example, not a lexical entry of RO):

- (9) *ilt- og* *brintatomer*
 oxygen- and *hydrogen.atoms*
 ‘oxygen atoms and hydrogen atoms’

In contrast to (7)–(8), the hyphen in (9) does in fact substitute a part of a word, as the expanded phrase shows: *iltatomer og brintatomer* ('oxygen atoms and hydrogen atoms'). In other words, it is a locum hyphen. Furthermore, the two conjuncts can be switched around, and the conjunction can be replaced with another conjunction.

There is thus a sharp contrast between on the one hand (7)–(8) and on the other hand (9). The hyphens in (7)–(8) are **not** real locum hyphens: they do not substitute a part of a word but only appear to do so. They are **pseudo** locum hyphens.

We thus make the distinction between locum hyphens and pseudo locum hyphens, and correspondingly between locum hyphen phrases and pseudo locum hyphen phrases. There might, however, be several types of pseudo locum hyphen phrases. See (10) (from §57.2):

- (10) de seks- til syvårige
 the six- to seven.year olds
 'the six to seven year olds'

The expanded phrase is not synonymous with (10) and is hardly meaningful at all (??*De seksårige til syvårige*, 'the six year olds to seven year olds'). Thus, the hyphen is not a locum hyphen but a pseudo locum hyphen.

However, in contrast to (7) and (8), the conjunction in (10) may be replaced with another conjunction (*og* 'and' or *eller* 'or'), giving rise to a well-formed semantic expression (although a different one from that of (10)). Given the differences in replaceability, we may correspondingly distinguish between different types of pseudo locum hyphen phrases – or at least note that they have different properties.

In any case, the pseudo locum hyphens in (7), (8) and (10) all seem to have a linking (or gluing) function: They link complex compounds consisting of more than one text word. Thus, the relevant article of RO would instead seem to be §57.7a and §57.7.b (here reformulated and simplified):

When the first part of a compound or a derivative consists of more than one word, a hyphen is inserted between the first part and the second part.

Following this rule, then, the hyphen in (10) should instead be placed as in (11):

- (11) de seks til syv-årige
 the six to seven.year olds
 'the six to seven year olds'

The hyphen links the complex first part *seks til syv* with the second part *årige*. The same would seem to apply to the hyphens (7)–(8).

Finally, in some cases it is not clear whether a hyphen is a locum hyphen or a pseudo locum hyphen. The hyphen in (12) is ambiguous between the two, as shown in the two translations:

- (12) salg.s- og leveringsbetingelser
sell.LINK- and delivery.conditions
'conditions for sale and delivery'
'conditions for sale and conditions for delivery'

The complicated cases in (3)–(5), the pseudo locum hyphens in (7), (8) and (10) and ambiguous cases such as (12) suffice to show what kind of challenges a parser faces when analyzing MWEs involving hyphens. In Section 4, we present a formal analysis to tackle some of these challenges.

4. The Formal Analysis of MWEs Involving Hyphens

COR₁ uses the CLINK lexical template (Henrichsen, 2023; 2024a), based on categorial grammar (the Lambek calculus). Readers unfamiliar with the “slash-style grammar” will find Wood (1993) and Henriksen (2024b) useful guides.

The basic data structure in categorial grammar is the so-called sequent, having the form ANTECEDENT ==> CONSEQUENT, where ANTECEDENT is a list of categories and CONSEQUENT is the hypothetical PoS of the phrase to be proven. Each category containing a slash ($y \backslash x$ or x / y) requires an argument y either to its left side ($y \backslash x$) or to its right (x / y) in order to produce a resulting category x . Table 2 provides an example.

Table 2: The CLINK lexical template

gas- og vandmestrene ('the-plumbers _{pl.def} ') [gas][-] 110 110\119 ==> 119 [og] X\X/X ==> X\X/X [vand][][mestrene] 119 Y/Z/Z 113 ==> 113		
119 X\X/X 119 Y/Z/Z 113		==> <i>Consequent</i>

The main sequent (the formula under the line) is satisfied for $X=119$, $Y=119$, $Z=113$, and *Consequent*=113, proving that *gas- og vandmestrene* ('the gas- and water.repairmen'; 'the plumbers') has the syntactic function of a noun in *uter.plur.def* (corresponding to the inflexion index 113). The proof also has a semantic projection. The COR₁ homepage at ordregister.dk provides an interactive PoS table relating the category indices of COR₁ (such as 110, 113, 119) and the inflexional forms of Danish morphology.

As a special feature of the categorial grammar analysis, each proof step (i.e., each application of a slash category with an adjacent category) is coordinated with a semantic composition, see details in Henriksen, 2024a). As a result, compositional

ambiguities appear in the form of alternative lambda-formulae, as exemplified in Table 3. Categories and semantic proxies are imported from COR₁ (thus *mestre* has the COR-id COR.51525.112).

Table 3: Lexical information for three COR₁ lexemes

Form	Category	Semantic proxy
<i>gas</i>	110	62857
<i>vand</i>	120	41032
<i>mestre</i>	112	51525

Alternative semantic projections of *gas- og vandmestre*:

51525(62857 & 41032)

51525(62857) & 51525(41032)

These two semantic analyses (simplified here for readability) correspond to the two paraphrases (i) ‘repairmen of gas and water’ or (ii) ‘repairmen of gas and repairmen of water’. These are obviously two different interpretations, as can be appreciated by the singular form *gas- og vandmester*. As mentioned, there is no such thing as a repairman of gas nor a repairman of water; a repairman of gas and water (i.e., a plumber), however, exists. It is not surprising that the singular phrase frequently occurs in ordinary Danish texts (actually even more frequently than the plural). In contrast, *ilt- og brintatomer* (‘oxygen and hydrogen atoms’) practically never occurs in the singular, as opposed to the plural, as indeed predicted by the semantic analysis. How could a single atom be of the class oxygen **as well as** of the class hydrogen?

5. Discussion

Our study has revealed a hitherto undescribed (or at least unpublished) classification of the MWEs with hyphens as they appear in ordinary Danish text. As shown, these constructions show considerably more complex semantic patterns than implied by their orthographic forms.

For instance, MWEs like *gas- og vandmester* may have non-compositional semantics, making them vulnerable to e.g., misguided translations (into e.g., ‘repairman of gas and repairman of water’, in place of ‘repairman of gas and water’).

Inspired by the German typology of hyphens (see for instance Fleischer & Barz, 1995, pp. 92, 122–125, 131–135, 142–143), we suggest that a Danish typology be developed. For instance, the function of the locum hyphen is very different from that of the hyphen in *B-skat* (‘B-tax’), which links *B* and *skat*. We consider the present paper as a step in that direction.

We are currently working on the proper inclusion in COR₁ of the locum hyphen and the pseudo locum hyphen, which in turn will facilitate the development of “intelligent”

text processing, including grammar checking and machine translation. As a valuable side effect of our formal approach, we may be able to suggest revisions to the RO rule articles with regard to hyphenation.

6. Summary

In this paper, we have discussed MWEs with hyphens, both in relation to the relevant rules and lemmas of Retskrivningsordbogen and in relation to some of the related challenges of NLP (natural language processing). Importantly, we have pointed to the hitherto unnoticed problem of the pseudo locum hyphen.

We have briefly presented a formal analysis based on categorial grammar and the lambda-calculus and shown how it can represent the semantics of MWEs with hyphens and their potential ambiguities.

Generally, formalization projects like the one reported here have often revealed indeterminacies in Retskrivningsordbogen, such as those related to the (pseudo) locum hyphen. Using formal methods as those employed in this paper makes it possible to pinpoint schisms between normative orthographic rules and the morpho-semantic reality.

References

- Act on Danish Orthography* (1997). Retsinformation, LOV nr. 332 14/05/1997.
- Bauer, L. J. (1975). *Nominal compounds in Danish, English and French*. [Doctoral dissertation, University of Edinburgh].
- Fleischer W., & Barz, I. (1995). *Wortbildung der deutschen Gegenwartssprache*, 2nd edition. Max Niemeyer.
- Henrichsen, P. J. (2023). Det Centrale Ordregister. Et indeks over det danske sprog – en gave til dansk sprogteknologi. In L. Holmer, G. Horn, H. Landqvist, P. Nilsson, E. Nordgren, & E. Sköldberg (Eds.), *Nordiska Studier i Lexikografi 16* (pp. 113–126) Nordiska Forening i Lexikografi, Lund Universitet.
- Henrichsen, P. J. (2024a). Make each morph count. A new approach to computational lexicography for text processing. In K. Š. Despot, A. Ostroški Anić, & I. Brač, *Proceedings of the 21st EURALEX International Congress, Lexicography and Semantics* (this issue).
- Henrichsen, P. J. (2024b). Tekstordet som grammatisk domæne. *Ny forskning i grammatik*, 31, 53–70.
- Jervelund, A. Å., Schack, J., Jensen, J. N., & Andersen, M. H. (2012). *Retskrivningsordbogen*, 4th edition. Alinea.

Widmann, T. (2024). The Central Word Register of the Danish Language. In M. Medved, M. Měchura, C. Tiberius, I. Kosem, J. Kallas, M. Jakubíček, & S. Krek (Eds.), *Electronic lexicography in the 21st century (eLex 2023). Proceedings of the eLex 2023 conference* (pp. 91–103). Lexical Computing CZ.

Wood, M. M. (1993). *Categorial Grammars*. Routledge.

Contact information

Michael Nguyen

The Danish Language Council
mn@dsn.dk

Peter Juel Henriksen

The Danish Language Council
pjh@dsn.dk

Pär Nilsson

FIGURATIVE, TRANSFERRED OR EXTENDED USE? The Use of Semantic Labels in the First Edition and the Revised Version of the Swedish Academy Dictionary

Abstract In this paper, a doctoral study on the description of semantic change in the Swedish Academy Dictionary (SAOB) is presented. The starting point for the study is semantic labels like figurative and in extended use. Five such labels in SAOB are examined, mainly with methods from the cognitive linguistic framework. The results show, among other things, that the most labelled mechanism is metaphor, that metaphor and metonymy often co-occur and that such co-occurrences often are expressed and explained in the dictionary with combined or modified labels. Furthermore, there seems to be a certain overlap between some of the labels in use, and one could question if all five labels are needed. Since the first volume of SAOB was completed in 2023 and a revision project now is launched, it is possible to make practical use of the results of the study in the dictionary. Therefore, some practical outcomes and applications that are either planned or already started are also presented in this paper.

Keywords lexicography; figurative language; semantic change; cognitive linguistics; conceptual metaphor theory; Swedish Academy Dictionary

1. Introduction

Semantic labels such as *figurative*, *transferred* and *extended use* are well established in lexicographical meta-language and occur in many dictionaries. In the extensive *Swedish Academy Dictionary* (SAOB 1893–) they are highly frequent; to give an example, the term *bildlig* ('figurative') alone is used more than 40,000 times. However, how these labels are used in practice in the dictionary and what the users can expect from them is not quite clear. It is not distinctly stated in the readers' manuals nor in the internal editorial handbooks. This is the starting point for Nilsson, 2023 who studies different types of mechanisms of meaning change, and the description of these in SAOB based on five such definition labels. The labels in question are *bildlig* ('figurative'), *oegentlig* ('un-actual/non-factual'), *överförd* ('transferred'), *utvidgad* ('extended') and *allmänare användning* ('more general use'). These are examined with different methods and from different perspectives, mainly within the framework of cognitive linguistics.

The question posed is how the labels are used in the dictionary and how systematic and consequent the use is. The aim is to find out what the labels in question mean in practice: what is a figurative use in SAOB? What is the difference between an extended and a more general use, and what, more precisely does it mean in the dictionary that a meaning could be un-actual? The purpose of the survey is to assess how relevant and consistent the analysis and the description method applied in SAOB (a dictionary in the making

for more than 130 years) are today, from a theoretical point of view. To what extent is the lexicographical description of the mechanisms behind semantic change compatible with modern semantic theory and in particular the ideas of cognitive semantics? From the opposite perspective, the purpose is to find out what the definitions in SAOB can teach us about semantic change and meaning-changing mechanisms. In addition, an overarching aim of the project is to make use of the results in the forthcoming revision of the dictionary. As from the beginning of 2024 a revision project is initiated, and thus, it is now possible to implement changes based on the findings from the study.

After this introduction a short description of SAOB as a dictionary follows in section 2 below. The survey, some of its different sub-studies and main results are presented in section 3. In section 4 practical applications of the results in the revision of the dictionary is presented, and in section 5 the discussion is summed up.

2. SAOB – A Short Introduction

SAOB is an extensive diachronic-contemporary dictionary that consist of 39 printed volumes and a corresponding digital version. The first volume was published in 1898, and the work was completed when the last volume was published and officially launched at the website saob.se in December 2023, thus more than 120 years later. The dictionary describes the Swedish language from the 1520s to the present day, the so-called modern Swedish period.

The microstructure of the dictionary entries is chronologically organized, with the most basic and primary meanings of the words presented before later developed and lexicologically secondary ones, as far as this is possible to establish. The semantic relations between the meanings and uses are also often described in much detail as can be seen in the entry ELEFANT ('elephant') in fig. 1 below (cf. also e.g., Holm & Jonsson, 1990; Eaker, 2006).

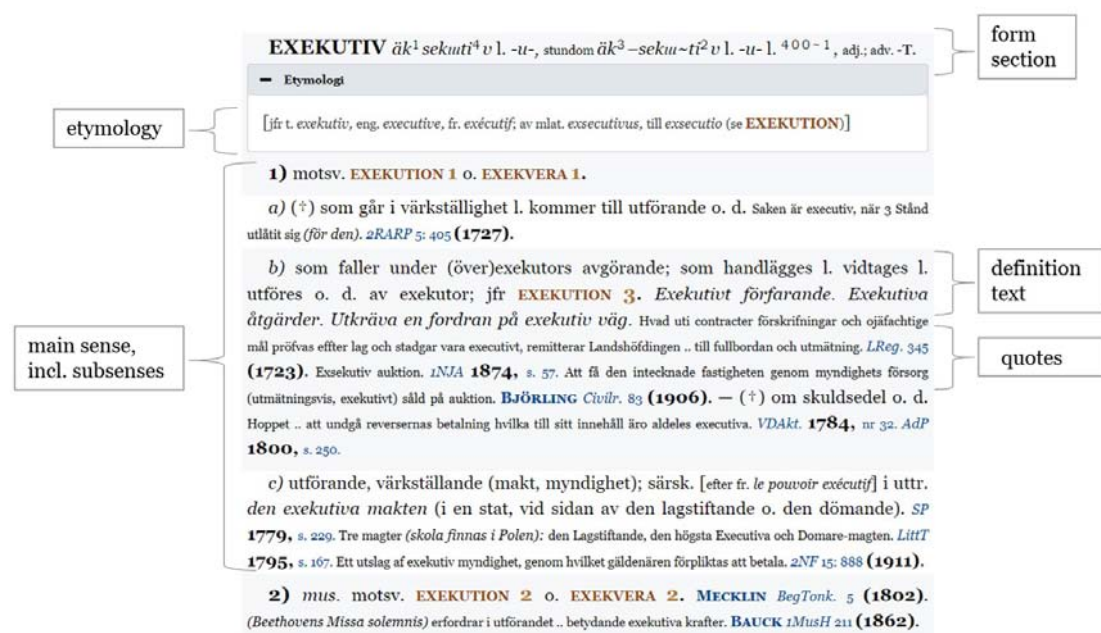


Fig. 1: Demonstration of SAOB's microstructure

The figure displays the principle of SAOB to mark main senses with Arabic numbers. In the entry the relation between the main sense of 1 and the secondary sub-sense (for clarity marked with a red box) is explained with the label “mer l. mindre bildl.” (‘more or less figurative’) and with an supplementing comment highlighting the elephant’s strength and clumsy physique, way of moving etc. This is a typical way of describing figurative extensions in SAOB.

Between the late 19th century and 2023 the work of SAOB was progressing alphabetically forward, without any revision of older parts. As mentioned above, the first edition is now completed, and due to the long publication time there is a great need for updating and revision. A revision project was initiated in the beginning of 2024 with grants that are estimated to last for about 7 years of work. The project aims at adding new entries, carrying out systematic changes in all or several entries at the same time, and, of course, updating old and outdated descriptions more or less manually. (For a somewhat fuller description of the dictionary, the history behind it, and the different parts of the revision project see Nilsson, 2024.) A particular type of descriptions in need of updating are those concerning semantic relations between senses.

3. The Study – Methods, Theoretic Framework, and Some Results

Common to those semantic labels that are examined is that they all describe meaning change/sense development, and the five most frequent such labels in SAOB were selected for the study. An additional selection criterion was that the labels were to denote denotative and semasiological change (in the sense of Geeraerts, 2015). In this context the term denotative refers to referential meaning (which can be contrasted with e.g., purely emotional meaning). Semasiological refers to the perspective of the single, individual word and the various concepts it denotes (as opposed to the term onomasiological, which instead focuses on the abstract concept and what words can be used to express this).

As mentioned above the label *bildlig* (‘figurative’) occurs over 40,000 times in SAOB. The label *oegentlig* (‘un-actual/non-factual’) is found approx. 11,000 times, *utvidgad* (‘extended’) about 10,000, *allmännare* (‘more general’) just under 8,000 and *överförd* (‘transferred’) a total of about 5,000 times. These labels are traditional in the sense that they are not coined by the SAOB editors – they (or their equivalents in other languages) appear in several others dictionaries, and often the SAOB editors seem to have been inspired by them. For example, expressions such as *transferred* and *extended use* are found in the Oxford English Dictionary (OED) and *bildlich* and *uneigentlich* in Deutsches Wörterbuch von Jacob Grimm und Wilhelm Grimm (DWB), both of which dictionary projects were initiated before the work on SAOB started. Concepts such as figurativeness and transfer are also discussed in the semantic literature.

The examined material consists of 300 occurrences of each label, hence 1,500 in total, extracted from the digital version of the dictionary, by way of free text search. In order to achieve a somewhat even distribution within the different parts (and between the years of production) of the dictionary, each amount of 300 was divided into four groups. Hence, 75 search matches per label were collected from each group.

Five different sub-studies were carried out for each semantic label, all of which were based on a certain research question in connection to semantic theory, especially cognitive linguistics and conceptual metaphor theory. In this paper three of the studies are presented in short.

With (mainly) the three mechanisms metaphor, metonymy and generalization as a benchmark, the first study examines *what semantic processes do the different labels denote?* The analysis becomes an attempt at matching to what extent for instance a use labelled figurative in SAOB corresponds to a metaphor, and an *allmänare*, ‘a more general use’, to a generalization. The method is based on Lakoff & Johnson’s (1980) definition of metaphor and metonymy: metaphors should not only be considered as a phenomenon in language, but as something related to our underlying way of think and act. Metaphors involve mapping between two different conceptual domains; and a domain is to be understood as a structured whole, a coherent organization of experiences. Metonymy means mapping within one and the same domain, or one and the same domain complex (cf. Croft, 1993), and a simultaneous shift of reference. The concept of generalization is, for various reasons, not as frequently discussed in conceptual metaphor theory. But what is usually mentioned in the discussion is that it is possible to question whether it is a separate process or just the result of some other process: if generalization is to be understood only as extension or expansion of a meaning, or a category of a meaning, metaphor becomes a subtype of generalization (cf. Bybee et al., 1994). In this study however, generalization and semantic extension is seen as two separate phenomena (in accordance with Koch, 2016). As far as domain mapping can be explained as a phenomenon of degree, generalization should belong to the same scale as metaphor. In this process, too, it is a matter of analogy between one source and one target concept; but for the generalization part it is not a question of mapping, or certain correspondences, across different domains, but within one and the same domain. The difference between generalization and metonymy lies in the fact that the latter process also requires a change of reference. However, both the result of a generalization and a metaphor can be seen as some sort of *semantic extension*.¹

The next study is concerned with the matter of (mapping between) concrete and abstract concepts and domains. The research question to be answered is based on the idea that often is expressed within cognitive linguistics (see e.g., Kövecses, 2016), namely that *a metaphor is typically constituted by a projection from a concrete to a more abstract domain – does this also apply in SAOB?* Hence, the task was to examine to what extent the derived labelled meanings in the material are more abstract than their basic meanings, or if the level of abstraction can be considered the same for the different concepts that are mapped together. A concrete concept in the study is defined as a sensory one, one that is physically perceivable (in accordance with Svanlund, 2001). In the studies mentioned at this point, not only the labels as such (and their corresponding basic meanings) are analysed, but also the (other) content of the current definitions and the quotations that illustrate the meanings in the dictionary.

¹ Due to the lack of space it is not possible to describe the detailed method in this paper, for a fuller description of the analysis template see Nilsson (2023, p. 116 ff).

In the last study it is examined to what extent the relevant labels are combined with other labels, or modified in other ways (as in Figure 1 above, where the derived meaning was labelled “*more or less figurative*”) and, also, what such modifications and combinations denote. In the cognitivist literature, the boundaries between metaphor and metonymy are described as fuzzy: a word or an expression can be used more or less metaphorically or metonymically; different mechanisms can also interact in certain ways (cf. Goossens, 1990, Croft, 1993, Ruíz de Mendoza Ibañez, 2000). The current sub-study is a way to study grey areas between different semantic processes. The questions posed in this respect are: *What labels are combined/modified and how? In what way do the combined/modified labelled uses differ from uses marked with a single label?*

The results show, among other things, that metaphor clearly is the most labelled process in the material. However, each label involves a certain amount of all three mechanisms metaphor, metonymy and generalization. In many cases it is a matter of interplay between the different mechanisms (à la Goossens concept of *metaphonymy*).

Almost every use labelled *bildlig* (‘figurative’) involves a metaphor in some respect. Other mechanisms involved for this label are mainly made up of metonymies. Also the label *oegentlig* (‘un-actual/non-factual’) is mainly represented by metaphors, but often of a different kind than the ones labelled *bildlig*. As a matter of fact, the label *oegentlig* often is an indication of what could be described as a less figurative metaphor, in terms of a less abstract target concept, than the metaphors labelled as *bildlig*. Typically a metaphor labelled *oegentlig* is represented by concepts concerning people (or animals or plants). An example from the material is the noun *taliban* and the metaphor INTOLERANT AND STRICT PERSON IS A FUNDAMENTALIST ISLAMIST, in quotes like the following, here translated one: “The talibans called for mandatory alcohol lock in all vehicles [...] and more speed controls”. In such cases the target domain and its projected concepts are not considered to be more abstract than the source.

However, the label *oegentlig* (‘un-actual/non-factual’) is also frequently used to denote generalizations, in particular, those involving some form of norm conflict, typically an expansion of a sense depending on a technically incorrect use of a technical term in general language. An example is the use of the word *mal* (‘clothes moth’) to denote any moth (although no attack on fabrics occurs, i.e., there is no actual “milling”, as the term originally refers to).

Also the label *överförd* (‘transferred’) involves a lot of metaphors. In the first half of the dictionary it seems to be used more or less equivalent to the label *bildlig* (‘figurative’). But as frequent is it used to denote a certain kind of metonymy, illustrated by the example: *(She wrote) an angry letter*, i.e., a case where characteristics from the person are transferred to the artefact related to the person. This use is found mainly in the second half of the dictionary. This shift of use has also been regulated in the internal editorial handbooks, since at least the 1990s.

Both labels *allmännare* (‘more general’) and *utvidgad användning* (‘extended use’) are dominated by the mechanism generalization. However they quite often also

represent metaphors. In such cases it is often either a question of generalizations that are combined with or difficult to distinguish from metaphors (see example below), or else the labels are used in regard to the result, as a notion overarching figurativeness, cf. expressions like: “in more general, normally figurative, use”, found in the material.

Combined or modified labels are very frequent in SAOB and the tendency is increasing over time. Quite often it is a matter of long and tortuous combinations (such as *in improper, extended or more or less figurative use*) that are difficult for the user to decipher. *Oegentlig* (‘un-actual/non-factual’) is the most combined/modified label – almost half of the labels studied are used in this way. Furthermore, there seems often to be legitimate motives for this kind of labelling (both regarding *oegentlig* and the other labels). Normally, it is a matter of semantic changes that involve more than one mechanism at the same time, such as metaphonymy, as mentioned above, but also combinations or interactions between metaphor and generalization are found – for example in expressions like *take damage/a hit*, where the sense ‘receive’ of the verb *take* (Swedish *ta*) has acquired a more general meaning ‘to be exposed to, affected by’; the expression as a complex can then be interpreted literally (in the case of physical violence) but also metaphorically (in the case of more abstract damage or loss etc.). In addition also combinations of generalization and metonymy are found in the material.

In sum, SAOB’s description of sense development by the means of semantic labels is nuanced and overall systematic. It put to foreground the grey areas between different mechanisms and the fact that they often are combined in practice. On the other hand, the descriptions are often complicated and difficult for the user to interpret. There is also a certain overlap between different labels that is uncalled-for. The overlap means that the user cannot be confident which semantic process is actually involved in a specific description. A conclusion is that the system used has clear points and could be worth keeping, but it has potential for improvements.

4. Practical Implications of the Study

Since 2024 SAOB is being revised. The description of semantic relations between meanings and uses is one defined assignment within the revision project. The purpose is to increase consistency of, and clarify, the use of labels, and to reduce overlap between them. A new order has been introduced, which implies that of the five abovementioned labels only three is still being valid, namely *bildlig* (‘figurative’), *utvidgad* (‘extended’) and *allmännare* (‘more general use’). In practice, this means that *oegentlig* (‘un-actual/non-factual’) and *överförd* (‘transferred’) are no longer allowed in new entries. In entries published in the first volume of the dictionary, the labels will be changed. These changes are (at the time of writing being initiated, and are) operated in a semi-automatic way, by identifying current occurrences and replacing with standard phrases. In the first half of the dictionary (approx. between the initial letters A–L) the label *överförd* can often easily be replaced by *bildlig*. In the second half (M–Ö) the metonymic use (“an angry letter” etc.) is replaced by phrases like: *of things or phenomena that are characterized by or bear witness of such (anger etc.)*.

SAOB's label *in metonymic use*, not mentioned to this point, and not a part of the study because of its very low frequency, is encouraged to be used more often if needed. Hence the three mechanisms of change, metaphor, metonymy and generalization, are closer connected to a single label in the dictionary. One must though bear in mind that the different labels to a different degree is a part of the general language vocabulary (cf. Nilsson & Laitinen, manuscript in preparation). Labels like *figurative* has a longer tradition in dictionaries, and probably many users have a more intuitive idea of the implications of such a process and of metaphors than of metonymies. On the other hand, the target group of SAOB is not only general language users, but also more advanced users. In long term, a possible solution to the problem would be to introduce a more detailed, advanced view.

In addition, the scope of the labels *utvidgad* ('extended') and *allmännare* ('more general') will be kept apart more clearly. As also stated in the internal handbook (but not always practised), the former label should be used to describe a (non-figurative) *development* towards a semantic extension, while the latter rather describes the extended *result* or put to focus a previously more general sense in relation to a later developed one.

Lastly, long and complicated combinations of labels will be avoided in the revised version of the dictionary. For the sake of the users' interpretability it is no longer allowed to combine more than two different labels.

5. Concluding Summary

To summarize, a study on the use of semantic labels in SAOB has been conducted. The results indicate a nuanced system of description that still have room for improvements. In the newly started revision project, the description of semantic relations constitutes one of the defined revision assignments. The main purpose of this part of the project is to avoid overlap of the labels and to make the use more consistent and more easily interpreted.

One could of course question the need for labels at all – wouldn't explicit explanations (in full sentences) be more than enough? Are the definitions not sufficiently detailed without the labels? The answer to these questions is yes, the user would probably often get roughly equivalent information from each entry even without the defining labels. However, often the description in SAOB consists only of the label, without any supplementary definition text, and these cannot simply be deleted without consequences. But more fundamental is the fact that the labels can contribute with information on the onomasiological level. Description of systematically recurring patterns of meaning change with recurring, precise and general terms implies, firstly, that the individual word (and its derived meanings) is related to the linguistic system and explained in accordance to this, and, secondly, that a user who is interested in changes concerning the conceptual system, by means of the semantic labels, gets access to tools for analysing (and practically searching) semantic patterns. From this point of view, the SAOB editors are looking forward to improving the dictionary and increase the areas of use.

References

- Bybee, J. L., Perkins, R. D., & Pagliuca, W. (1994). *Evolution of grammar: tense, aspect, and modality in the languages of the world*. Chicago: Univ. of Chicago Press.
- Croft, W. (1993). The role of domains in the interpretation of metaphors and metonymies. *Cognitive linguistics*, 4, 335–370.
- DWB: *Deutsches Wörterbuch*, von Jacob Grimm und Wilhelm Grimm, hrsg. von der Deutschen Akademie der Wissenschaften zu Berlin, Band 1–16 + Quellenverzeichnis, Leipzig 1852–1971. In: <<https://woerterbuchnetz.de/?sigle=DWB#0>> (last access: 30 May 2024).
- Eaker, B. (2006). Historisk lexikografi i praktiken. Om momentordningen i SAOB. *LexikoNordica*, 13, 41–54.
- Geeraerts, D. (2015). How words and vocabularies change. In J. Taylor (Ed.), *The Oxford Handbook of the Word* (pp. 416–430). Oxford: Oxford University Press.
- Goossens, L. (1990). Metaphonymy: the interaction of metaphor and metonymy in expressions for linguistic action. *Cognitive Linguistics*, 1(3), 323–340.
- Holm, L., & Jonsson, H. (1989–1991). Swedish lexicography. In F. J. Hausmann (Ed.), *Wörterbücher: ein internationales Handbuch zur Lexikographie = Dictionaries: an international encyclopedia of lexicography* (pp. 1933–1943). Berlin: de Gruyter.
- Koch, P. (2016). Meaning change and semantic shifts. In P. Juvonen, & Koptjevskaja-Tamm, M. (Eds.) *The lexical typology of semantic shifts* (pp. 21–66). Berlin: De Gruyter Mouton.
- Kövecses, Z. (2016). Conceptual metaphor theory. In E. Semino, & Zs. Demjén (Eds.), *The Routledge Handbook of Metaphor and Language* (pp. 13–27). Milton Park, Abingdon, Oxon: Routledge.
- Lakoff, G., & Johnson, M. (1980). *Metaphors We Live By*. The University of Chicago Press.
- Nilsson, P. (2023). *Lexikal betydelseutveckling i teori och praktik: en analys av fem definitionsformler i Svenska Akademiens ordbok och de semantiska förändringsmekanismer som de beskriver*. Ph.D. thesis. Lund: Lund University.
- Nilsson, P. (2024). Report on the revision of the Swedish Academy Dictionary – and the search for “old neologisms”. In K. Š. Despot, A. Ostroški Anić, & I. Brač, *Proceedings of the 21st EURALEX International Congress, Lexicography and Semantics* (pp. 482–497). Institute for the Croatian Language.
- Nilsson, P., & Laitinen, S. (Manuscript in preparation). Hanteringen av metonymier i några en- och flerspråkiga nordiska ordböcker. In *Nordiska studier i lexikografi*, 17. Nordiska föreningen för lexikografi.
- OED: *The Oxford English Dictionary*, second edition, vol. I–XX (of the corrected reissue (1933) with an Introduction, Supplement, and Bibliography of A New English Dictionary on Historical Principles, vol. I–X, ed. by James A. H. Murray et al. (1884–1928) and Supplement, vol. I–IV, ed. by R. W. Burchfield (1972–1986)), Oxford 1989. Retrieved 30 May, 2024, from www.oed.com.

Ruiz de Mendoza Ibáñez, F. J. (2000). The role of mappings and domains in understanding metonymy. In A. Barcelona (Ed.), *Metaphor and Metonymy at the Crossroads. A Cognitive Perspective* (pp. 109–132). Mouton.

SAOB: *Svenska Akademiens ordbok* (1898–2023). Lund: Gleerups. Retrieved 30 May, 2024, from www.saob.se and www.svenska.se

Svanlund, J. (2001). *Metaforen som konvention. Graden av bildlighet i svenskans vikt och tyngdmetaforer*. Ph.D. thesis. Stockholm: Almqvist och Wiksell.

Contact information

Pär Nilsson

Deputy editor-in-chief, the Swedish Academy Dictionary
Ph.D., affiliated to Lund University

Svenska Akademiens ordboksredaktion
Dalbyvägen 3
SE-22460 Lund
Sweden

par.nilsson@svenskaakademien.se
par.nilsson@nordlund.lu.se

Sanni Nimb, Ida Flörke, Sussi Olsen, Bolette S. Pedersen, and
Nathalie C. H. Sørensen

COR.SEM, A NEW FORMAL SEMANTIC LEXICON FOR DANISH

Abstract We present the COR.SEM lexicon, an open-source semantic lexicon for general AI purposes funded by the Danish Agency for Digitisation as part of an AI initiative embarked upon by the Danish Government in 2020. COR.SEM describes the core senses of 34,000 Danish lemmas with formal semantic information, e.g., ontological type, hypernym, semantic frame, regular polysemy pattern, and polarity value; features which are in essence drawn and simplified from other existing resources. Lexical information from The Danish Dictionary DDO and the Danish Thesaurus DDB is also integrated, e.g., user examples, domain label, synonyms, and near synonyms. It provides direct links to synsets in the Danish WordNet DanNet, as well as to the morphological lemma information in COR, the Central WordRegister which is based on the Danish Orthographical Dictionary and DDO. The register's common numerical index at both lemma and sense level makes it more straightforward to merge mono- as well as bilingual dictionaries with COR.SEM and thereby inherit the formal semantic information. At the website corsem.dsl.dk it is possible to browse the lexical entries and to download tailored extracts of data of your choice. We give examples of the use of COR.SEM in linguistic studies, in NLP tasks and in lexicographic projects.

Keywords semantic lexicon; lexical semantics; linked data; computational lexicography

1. Introduction

COR.SEM is a new semantic lexicon for Danish published in 2024 at ordregister.dk in the form of a dataset for free download (CC0 license). The dataset is linked to the indexed morphological lexicon COR which is based on RO (the Danish Orthographical Dictionary) and available at the same site. The Central WordRegister has been developed as a public state initiative with the overall aim of facilitating and easing the transition of Danish industry and public institutions towards an increased use of language technology and language-centric AI. The vision has been to support the development of the technology so that it can function seamlessly in a Danish societal context as well as in different professional settings. To this end, a central premise of the COR project is that the resource be based on existing high-quality dictionaries and thus should take advantage of the rich information sources already provided in and about Danish culture and society. In section 2, we describe the existing dictionaries and resources and how the COR.SEM was compiled based on linked data. The initiative has made it more straightforward to share and merge existing Danish lexical resources via the register's common numerical index, which ensures a homogeneous approach to all incoming resources, be they terminologically, historically or in other ways specific. This feature has in fact been requested from Danish companies and public institutions for a long time.

COR.SEM gives open access to several semantic information types in a standardized and manageable way. This includes information such as a coarse-grained sense inventory of Danish, sentiment, and ontological information on Danish concepts as well as frame semantic information on Danish verbs and deverbal nouns. In section 2 we describe the lexicon. In section 3 the webpage CORSEM.DSL.DK is presented, and in section 4 we discuss different types of use of the resource before we conclude in section 5.

2. The COR.SEM Lexicon: Content and Method

The lexicon contains 34,020 lemmas (28,725 nouns, 2770 verbs, 2343 adjectives, and 177 adverbs) from the Danish Dictionary DDO (Hjorth & Kristensen, 2003–2005; ordnet.dk/ddo) and the Danish WordNet DanNet (wordnet.dk/dannet; Pedersen et al., 2009). Via links to DanNet in COR.SEM, more than 5,000 English equivalents can be identified. 11,300 COR.SEM lemmas are central lemmas in Danish being defined as either linked to one of the 5,000 base concepts in Princeton WordNet or being a central word (a highlighted keyword) in the Danish thesaurus DDB (Nimb et al., 2014; Nimb et al., 2014/2015) – or both. In this way we ensured that both the most central lemmas as well as a large variety of themes and semantic fields are well represented. The information on centrality is part of the resource. The degree of centrality is represented as a number (3, 2 or 1), in contrast to the 24,000 non-central lemmas (0).

6,000 of the lemmas are polysemous having a total of 15,000 COR.SEM senses. The sense inventories of these lemmas have been established by carefully studying the senses in the fine-grained DDO dictionary, discarding rare and uncommon senses and lump cases of semantically close senses, typically in the cases of narrower subsenses of a main sense. In this way, we have reduced the number of senses by 25% (see Section 2.1. for more details).

The sense information is expressed in the form of formalised information types assigned as a value from a closed list: ontological type, hypernym, frame, and polarity value. Furthermore, synonyms from the DDO and DanNet (in the form of `synset_members` of the same `synset`), as well as a small group of semantically related words extracted automatically from the thesaurus DDB contribute to the description of the sense.

All senses in COR.SEM have at least one, sometimes two ontological type values (8% of the cases), and at least one, sometimes two hypernyms (8% of the cases). The ontological type is chosen among 146 values from an established COR.SEM ontology, a simplified version of the DanNet and EuroWordNet ontologies (Vossen, 1999).

The frame values correspond to the frames in Berkeley FrameNet, where they are further specified, and the polarity values follow a scale from minus to plus 3 (0-values are not marked).

The polarity values were transferred from working files from the sentiment lexicon project (*Det Danske Sentiment-leksikon*; Nimb et al., 2022). They were validated and equilibrated in the case of lumped DDO senses in COR.SEM. Also new senses were assigned a polarity value. In a supplementary resource at ordregister.dk, COR.SEM. EXT, definitions and examples from the DDO are available under a more restricted license (CC BY-NC-ND).

In Figure 1 the verb *håbe* ('to hope') is shown, illustrating the formal semantic information types. Its ontological type is 'Act+Mental', the semantic frames are 'Desiring' and 'Reliance_on_expectation'. The sentiment polarity value is +2 (on a scale from -3 to +3). Links to the equivalent synset and the hypernym in DanNet are also provided.

In Table 1 the numbers of lemmas, senses and values are shown.

Betydning 1: have et forventningsfuldt ønske om at noget bestemt...

Synonym fra DDO: nære håb

Emne: psy

Brugseksempler:

Overbegreb: ønske

Relaterede ord fra DDB: *formode, lukke øjnene og ønske, synes, ønske at*

Ontologisk Type: Act+Mental

Frame: Desiring, Reliance_on_expectation

Sentiment: 2.0

Se i DanNet:

DanNet synset: [synset-46847](#)

DanNet overbegreb: [synset-42934](#)

Fig. 1: The entry of the monosemous verb *håbe* ('to hope') with the (automatically shortened) DDO definition 'have an expectant desire for something specific', the DDO synonym *nære håb*, the DDO domain 'psy' (psychology), the hypernym *ønske* ('to wish'), related words from DDB ('suppose', 'close the eyes and wish', 'think', 'wish to'), the ontological type 'Act+Mental', the frames 'Desiring' and 'Reliance_on_expectation', and the positive sentiment value 2. Links to the corresponding synset in DanNet as well as the hypernym synset in DanNet are also given

Table 1: The numbers of information types for lemmas and senses in COR.SEM

Information type	Lemmas	Senses	Values, total number
Synonyms from DDO	11,557	13,236	19,959
Related words from thesaurus	32,832	38,830	114,834
Usage examples from DDO	25,275	32,979	46,993
Frame values (Berkeley FrameNet)	6720	8806	12,891
Systematic polysemy	1100	1609	1609
Sentiment value	5730	6823	6823

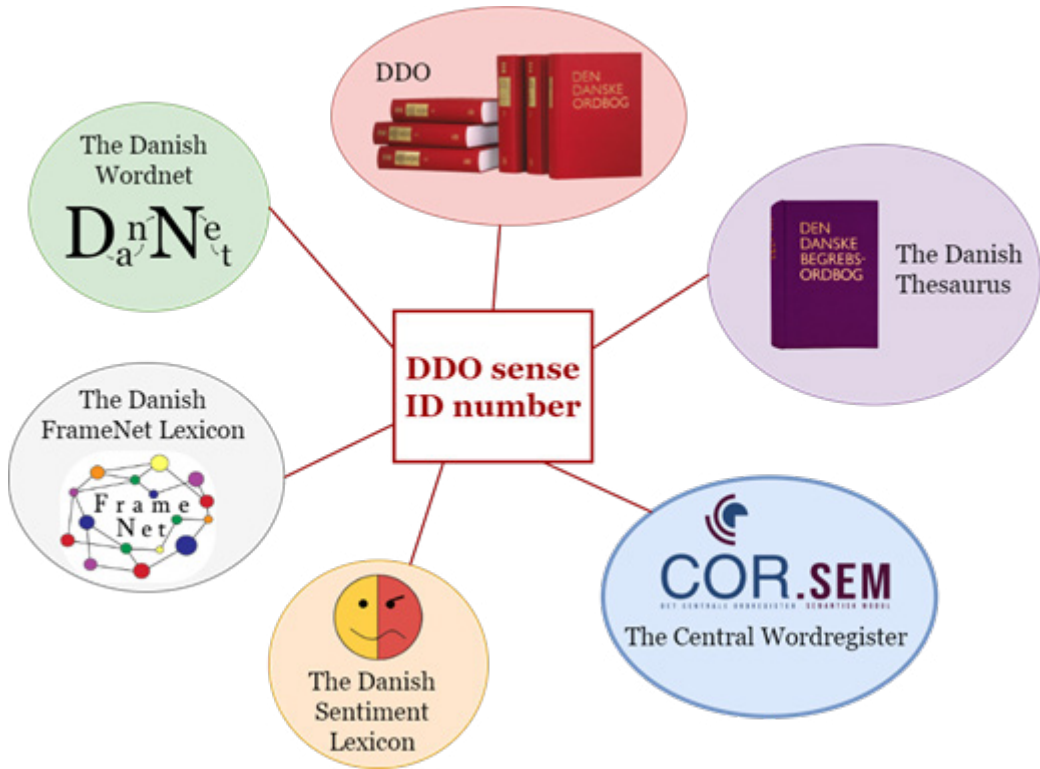


Fig. 2: The linked dictionaries and lexicons. Information from DDO, the thesaurus DDB, DanNet, The Danish FrameNet Lexicon and the Danish Sentiment Lexicon constitute the skeleton of COR.SEM

2.1 Linked Data

The COR.SEM lexicon was compiled based on already existing information in five dictionaries and formal semantic lexicons. In addition to the three mentioned above - DDO, DanNet, and the Danish thesaurus DDB - also The Danish FrameNet Lexicon (Nimb et al., 2017), and The Danish Sentiment Lexicon (Nimb et al., 2022) were used.

The five resources constitute a lexical network since they are all linked at sense level to the DDO sense inventory, see Figure 2. By keeping track of the shared sense id numbers, COR.SEM is today integrated in the network representing both semasiological and onomasiological information for approx. $\frac{1}{3}$ of DDO's lemmas and more than half of the lemmas in DanNet and the RO orthographic dictionary.

Opposite to DanNet, which in a ‘classical’ wordnet manner does not necessarily contain all senses for each lemma, COR.SEM covers all common senses of a lemma, except fixed expressions.

The fine-grained sense distinctions that we find in dictionaries for humans such as DDO, constitute a problem for language technology and AI. Experience from former projects, e.g., SemDaX, see Pedersen, B.S. et al. (2016), shows that coarse-grained sense distinctions work much better. For that reason, in COR.SEM we collapsed the fine-grained DDO senses of many polysemous words into fewer senses. We describe the principles followed in the next section.



Fig. 3: The DDO and COR.SEM entries of the adjective *elektrisk* ('electric'). In DDO to the left, the adjective has five senses, in COR.SEM to the right it has only two since three of the DDO subsenses are lumped with the main sense

2.2 Compiling a Core-Grained Sense Inventory Based on a Fine-Grained Monolingual Dictionary

The main principle for lumping senses was to keep the DDO main senses of a lemma and collapse a subsense with its main sense since subsenses in DDO are closely semantically related to their main sense. If, however, a subsense is of a different ontological type, which will typically be the case for figurative senses of a lemma, it will not be collapsed with its main sense. For instance, for the word *elektrisk* ('electric') in Figure 3, COR.SEM has two senses while DDO has five: one main and four subsenses. The DDO main sense and three of its four subsenses are merged into one sense (Betydning 1 ('Sense 1')). These senses all describe electricity in its physical sense: concerning/functioning with electricity, and they share the ontological type: Property_Physical. The last subsense is a figurative sense: 'characterised by tension, intensity' with the ontological type: Property+Mental. This subsense has its own sense in COR.SEM (Betydning 2 ('Sense 2')) in Figure 3).

For cases that are not straightforward, other factors influence the decision of whether to collapse a sense or not. We use calculated weight scores for each sense, based on the amount of extra lexicographic information connected to the sense in DDO, e.g., frequency information, collocations, and number of examples (no example in DDO is often due to low corpus frequency), to decide whether the sense is essential and consequentially should not be collapsed with other senses.

In Figure 4 the COR.SEM entry of the noun *afsløring* (‘disclosure’, ‘revelation’, ‘reveal’) illustrates how five senses in DDO ends up having only two senses in COR.SEM. Betydning 1 (Sense 1) in COR.SEM covers both the DDO main sense 1 (‘discovery of a (dishonorable) relationship that has been tried to be hidden’) and the DDO main sense 2 (‘an investigation and discussion of conditions that have hitherto been hidden or kept secret’) since they both reflect the act of discovering something secret or hidden and of disclosing the information to someone. A strong argument for the lumping is that the two senses share ontological types (Act+Mental and Act+Communication). The subsenses of the main DDO sense 2 (‘information about something previously kept secret’; ‘detection by means of a survey etc.’) can also be described by these two ontological types and are therefore included in the broad COR.SEM sense 1.

afsløring *sb., fl.*

Betydning 1: opdagelse af noget som hidtil har været skjult

Brugseksempler:

Overbegreb: handling

Relaterede ord fra DDB: *afdækning, oplysning, udtalelse*

Ontologisk Type: Act+Mental, Act+Communication

Frame: Reveal_secret

Systematisk Polysemi: PROCESS / RESULT (ABSTRACT)

Betydning 2: højtidelighed hvor en skulptur, et mindesmærke e.l...

Emne: kun

Brugseksempler:

Overbegreb: højtidelighed

Relaterede ord fra DDB: *fejring, indvielse, kulturformidling, åbning*

Ontologisk Type: Event+Social

Frame: Cause_to_perceive, Social_event

Fig. 4: In COR.SEM the noun *afsløring* (‘disclosure’, ‘revelation’, ‘reveal’) has only two senses because of the lumping of two main DDO senses as well as two DDO subsenses which all concern the revealing of information. The third main sense in DDO describes the unveiling of something concrete to an audience and is therefore kept separately. It constitutes sense 2 in COR.SEM.

Opposite, we consider the third DDO main sense ‘ceremony where a memorial or the like is unveiled’ (in DDO illustrated by the example ‘the small circle of invitees was to witness the unveiling of a commemorative plaque’) to be a different sense. It is therefore not lumped with sense 1, but inserted in COR.SEM as sense 2 (Betydning 2) with the ontological type Event+Social.

The FrameNet frames assigned to the two COR.SEM senses (as this is a verbal noun) consolidate the decision of merging the four DDO senses and keeping the last one apart. ‘Reveal_secret’ for the first COR.SEM sense and ‘Cause_to_perceive’ and ‘Social_event’ for the second sense correlate closely with the ontological types.

A specific problem regarding the lumping of senses regards the approach to systematic polysemy in DDO where polysemy sometimes results in two DDO senses, and sometimes in one, regardless the type. A set of 30 systematic polysemy patterns with specified guidelines on whether to lump or split the two senses ensures a homogenous treatment in COR.SEM (see Sørensen et al., 2023). The patterns are part of the information in the lexicon. Figure 5 has been generated from the visualisation function (‘datavisning’) at the website corsem.dsl.dk (see section 3), and it shows the distribution of systematic polysemy patterns for nouns and verbs. The two PROCESS / RESULT patterns, one which represents concrete results, where COR.SEM splits in two senses, and one which represents abstract results, where COR.SEM lumps the two senses into one, constitute more than half of the noun cases. Examples are *forventning* (‘expectation’) (two senses in DDO, one sense in COR.SEM), and *produktion* (‘production’) (two senses in DDO, also two senses in COR.SEM). Looking at the distribution for verbs, ACT / EVENT is by far the most common pattern. The (typically) two senses in DDO are in this case not lumped but both represented in COR.SEM.

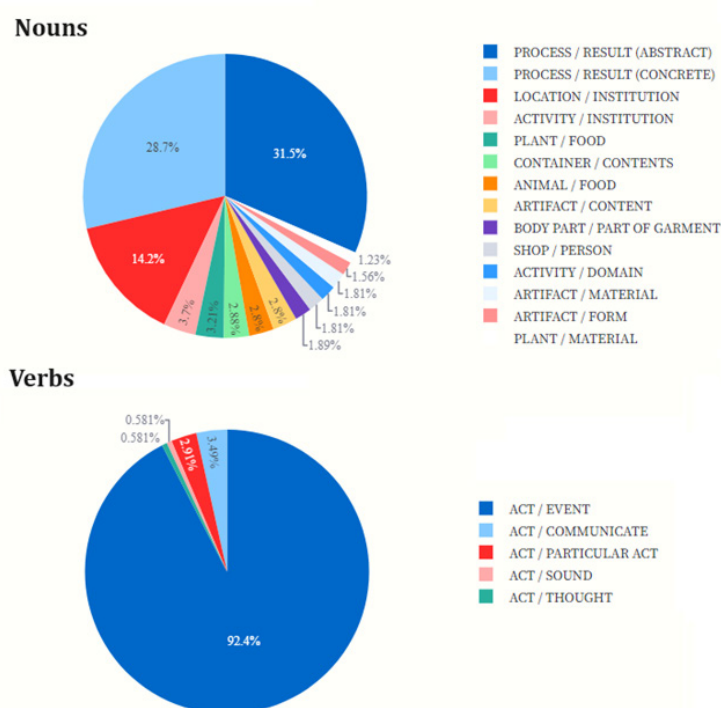


Fig. 5: The distribution of systematic polysemy patterns of nouns (top) and verbs (bottom) in COR.SEM

3. The Website CORSEM.DSL.DK

The cor.sem resource is published at ordregister.dk where it can be downloaded as a tsv-file¹. We have also created a Danish webpage **corsem.dsl.dk**² aimed at linguists and lexicographers, allowing them to view and interact with the Danish data, as well as to download either the entire dataset or selected parts of their own choice.

COR.SI	CO	DDO-opslagsord	bet	DD	ontologisk-type	synonym
COR.SI	abetræ	1	sb.	Plant+Object	Araucaria araucana	
COR.SI	aftenpragtstjerne	1	sb.	Plant+Object	Melandrium album pra	
COR.SI	agave	1	sb.	Plant+Object	Agave	
COR.SI	agerkål	1	sb.	Plant+Object	Brassica campestris	
COR.SI	agermåne	1	sb.	Plant+Object	Agrimonia eupatoria ali	
COR.SI	agertidsel	1	sb.	Plant+Object	Cirsium arvense	
COR.SI	akacie	1	sb.	Plant+Object	Acacia akacietræ	
COR.SI	akeleje	1	sb.	Plant+Object	Aquilegia	
COR.SI	alfagræs	1	sb.	Plant+Object	Stipa tenacissima engle	
COR.SI	aloe	1	sb.	Plant+Object	Aloe	

Download dataudsnittet (Licens: CC0)

ontologisk-type

Sorter efter værdiens frekvens i dataudsnittet Vælg antal som vises

Fig. 6: The website corsem.dsl.dk and the function 'Datavisning' ('demonstration of the data'). The user has chosen ontologisk type = 'Plant+Object', the wordclass 'sb.' ('noun') and 'synonym', the synonym from DDO, which is in this case often a Latin term

The webpage contains two types of viewing functions. The first one shows the content in line with a traditional dictionary presentation, presenting all semantic information for a selected lemma (see Figure 1, 3 and 4 above). This also includes definitions and (if activated) usage examples from the DDO. The second viewing function (see Figure 7) is aimed at giving the user an idea of the different types of semantic information across the entire COR.SEM vocabulary. It focuses on data and combinations of data of your own choice and allows the user to search for specific words and select information types to be shown on the screen. The data is presented as a table where each column represents one of the selected information types. As default, the data table contains a) an ID-column with an ID-string that is unique for a specific sense and is linkable to the lemma across the COR-resource, b) the lemma from the Danish Dictionary (DDO), c) the sense number in the resource, and the word class from the Danish Dictionary (DDO). The user can select additional columns to be shown in the data table. Additionally, we have created a filter functionality allowing the user to filter each column for specific values. With this functionality, it is for example possible to extract all negative words for animals (the filters ontologisk type = 'Animal+Object' and

¹ COR.SEM is open source (license CC0), however the DDO-definitions and examples are kept in a separate file COR.SEM.EXT under the license CC BY-NC-ND

² It should be noted that no formal usability evaluation of the website has been carried out

sentiment < 0). Another example is shown in Figure 6, where the user has selected all nouns with the ontological type ‘Plant+Object’ combined with the synonym from DDO, which is in this case is typically a Latin term³. It is also possible to generate diagrams showing the frequency of every unique information type in the selected data, as well as to download the data as tsv-files for further elaboration and application.

4. Making Use of COR.SEM Data

COR.SEM can be integrated as an independent semantic component in Danish language technology applications to improve aspects of Natural Language Understanding (NLU). Depending on the requested functionality this could for example be for sentiment analysis, for refined entity recognition, or for relation extraction. It can however also be used for training and/or fine-tuning of large language models (LLMs). By training and fine-tuning, the resource can help remedy some of the language transfer bias that are seen with current Danish LLMs where a substantial part of the content is in fact translated from English and thus not representing the Danish language, culture, and society in an appropriate way.

COR.SEM can also be used for benchmarking large language models (LLMs) for Danish. In Pedersen et al. (2024a) we demonstrate how it can be used as a benchmark for sense disambiguation. The coarse-grained sense inventory of COR.SEM combined with the usage examples (originally from DDO) were used to test whether the LLMs can disambiguate words in contexts. The usage examples of two senses of a lemma were mixed and put together in groups of two, some of which were indeed from the same sense, others which were not. The models are asked to identify whether the two examples represented two different senses of the lemma or only one and the same sense.

In another case of benchmarking, the sentiment value of a sense combined with a user example (manually confirmed to have the same polarity as the sense) is used to test whether the models can identify the polarity of the example (see Pedersen et al., 2024b).

Since the focus is on creating the evaluation datasets, only two models are tested, namely ChatGPT3.5 Turbo and ChatGPT4. In the word-in-context case, ChatGPT 4.0 performs best with 66% correct responses on the polysemous data. In the case of the sentiment datasets, ChatGPT 3.5 Turbo achieves a Pearson coefficient of 69.6%, while ChatGPT 4.0 achieves a coefficient of 81.9% (where the perfect result is a coefficient of 100%).

³ The Latin terms are not included as entries in DDO, only presented as synonyms to the Danish lemma sense.

Table 2: German verbs with Danish monosemous equivalents in a bilingual dictionary can inherit the formal semantic information in COR.SEM

German verb	equivalent COR.SEM monolingual verb	Frame (Berkeley FrameNet)	Ontological Type	Sentiment
<i>dementieren</i>	<i>dementere</i>	Respond_to_proposal Statement	Act+Communication	-1.0
<i>tagträumen</i>	<i>dagdrømme</i>	Cogitation Desiring	Act+Mental	
<i>blanchieren</i>	<i>blanchere</i>	Cooking_creation	Act+Physical+Purpose	
<i>bevormunden</i>	<i>beskænke</i>	Offering	Act+Possession	
<i>dezentralisieren</i>	<i>decentralisere</i>	Leadership	Act+Social	
<i>desillusionieren</i>	<i>desillusionere</i>	Stimulate_emotion	Cause+Mental	-2.0
<i>diskutieren</i>	<i>diskutere</i>	Discussion Judgment_communication	Act+Communication	

Within lexicographic work, COR.SEM data can be applied for many different purposes. It can be used to identify the words and meanings that are most relevant to include in a bilingual dictionary where Danish constitutes the source language. In the case where Danish is the target language, the information in COR.SEM on monosemy (constituting 28,000 lemmas) can be used to identify whether the Danish equivalent in the dictionary is monosemous. If this is the case, the formalised information in COR.SEM, e.g., ontological type, frame value and sentiment value can be automatically transferred to the source language. See Table 2 for examples from German/Danish. The plants in Figure 6 which contain a Latin synonym, is another example.

In monolingual lexicographic work, COR.SEM data can also be useful. At the Society for Danish Language and Literature, we are currently working on developing a DDOLite where we aim at using COR.SEM information as a baseline. The large amount of information in the fine-grained distinctions that we find in DDO can be an obstacle for students, learners of Danish and people with reading deficiencies who want to look up the meaning of a word. In these cases, the coarse-grained sense inventory of COR.SEM can be applied. In connection with the future DDOLite, which targets the above demographics, COR.SEM provides both a valuable data source and a starting point, especially regarding the cases where we condensed the fine-grained DDO senses of a lemmas into fewer senses. An example of this is the adjective *elektrisk* ('electric') in Figure 3. DDO has five senses while COR.SEM reduces this to two: 'regarding electricity' and 'characterised by tension, intensity'. A DDOLite could predictably be built on top of this information baseline, thus avoiding the complexity learners and people with reading deficiencies encounter when visiting DDO today. This new platform would only unfold a lemma in its central senses.

Finally, it should also be mentioned that COR.SEM can be used to study the Danish vocabulary in new ways. For instance, it is possible to gain new insights into the distribution of regular polysemy-patterns (see Figure 5), as well as to look for patterns that are not yet described in the literature. Likewise, combinations of information

types in COR.SEM often reveal interesting facts, e.g., that monosemous verbs are most likely to have the ontological type ‘Act + Communication’ and at the same time negative polarity. Some examples are *bagvaske* (‘to slander’), *bebrejde* (‘to blame’), *anklage* (‘to accuse’), *efterplapre* (‘to repeat what somebody says’, ‘to parrot’), and *formane* (‘to admonish’). The site CORSEM.DSL.dk facilitates this type of linguistic and lexical study for researchers, as well as for students and others with an interest in the Danish language.

5. Conclusion

COR.SEM constitutes a highly informative lexical semantic resource for Danish to be used in NLP as well as for lexicographic purposes. It also allows for new types of studies of the Danish language, facilitated by the strictly formalised semantic information types. The webpage corsem.dsl.dk presents the large amount of data and facilitates the studies and use of the content. We aim to expand the lexicon with more lemmas based on the vocabulary in the Danish orthographical dictionary RO which is updated once a year as well as bi-annual DDO updates. We also want to add frequent lemmas which are not yet covered, focusing on simplex words that are often part of composita, some of which are already included in the lexicon.

References

- Baker, C. F., Fillmore, Ch. J., & Cronin, B. (2003). The structure of the FrameNet database. *International Journal of Lexicography*, 16(3), 281–296.
- Berkeley FrameNet. Retrieved May 27, 2024 from <https://framenet.icsi.berkeley.edu/>
- DanNet. Retrieved May 27, 2024 from <https://wordnet.dk/DanNet>
- Den Danske Ordbog (DDO) (2009-). Online ordnet.dk/ddo. Det Danske Sprog- og Litteraturselskab.
- Det Danske FrameNet-leksikon. Retrieved May 27, 2024 from <https://korpus.dsl.dk/resources/details/framenet.html>
- Det Danske Sentiment-leksikon. Retrieved May 27, 2024 from <https://github.com/dsldk/danish-sentiment-lexicon>
- Fellbaum, Ch. (Ed.). (1998). *WordNet: An Electronic Lexical Database*. MIT press.
- Hjorth, E., & Kristensen, K. (Eds.). (2003–2005). *Den Danske Ordbog*. Det Danske Sprog- og Litteraturselskab og Gyldendal.
- Nimb, S., Braasch, A., Olsen, S., Pedersen, B. S., & Søgaard, A. (2017). From Thesaurus to FrameNet. In I. Kosem, J. Kallas, C. Tiberius, S. Krek, M. Jakubíček, & V. Baisa (Eds.), *Electronic Lexicography in the 21st century: Proceedings of eLex 2017 conference* (pp. 1–22). Lexical Computing Ltd.

Nimb, S., Lorentzen, H., Troelsgård T., & Theilgaard, L. (2015). *Den Danske Begrebsordbog*. Det Danske Sprog- og Litteraturselskab

Nimb, S., Olsen, S., Pedersen, B. S., & Troelsgaard, T. (2022). A Thesaurus-based Sentiment Lexicon for Danish: The Danish Sentiment Lexicon. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Language Resources and Evaluation Conference: LREC2022*, Vol. 2022 (pp. 2826–2832). Marseille. European Language Resources Association.

Nimb, S., Sørensen, N. H., & Troelsgård, T. (2018). From standalone thesaurus to integrated related words in the Danish Dictionary. In J. Čibej, V. Gorjanc, I. Kosem, & S. Krek (Eds.), *Proceedings from Euralex 2018* (pp. 915–923). Ljubljana, Slovenia. Znanstvena založba Filozofske fakultete Univerze v Ljubljani / Ljubljana University Press, Faculty of Arts

Nimb, S., Trap-Jensen, L., & Lorentzen H. (2014). The Danish Thesaurus: Problems and Perspectives. In A. Abel, C. Vettori & N. Ralli (Eds.), *Proceedings of the XVI EURALEX International Congress: The User in Focus* (pp. 191–199). EURAC research

Pedersen, B. S., Braasch, A., Johannsen, A. T., Martinez, H. A., Nimb, S., Olsen, S., Søgaard, A., & Sørensen, S. (2016). The SemDax Corpus. Sense Annotations with Scalable Sense Inventories. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Moreno | J. Odijk & S. Piperidis (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)* (pp. 842–847). Paris: European Language Resources Association. <https://aclanthology.org/L16-1136>.

Pedersen, B. S., Sørensen, N. C. H., Nimb, S., Flörke, I., Olsen, S., & Troelsgård, T. (2022). Compiling a Suitable Level of Sense Granularity in a Lexicon for AI Purposes: The Open Source COR-Lexicon. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Language Resources and Evaluation Conference: LREC2022*, Vol. 2022 (pp. 51–60). Marseille: European Language Resources Association. <https://aclanthology.org/2022.lrec-1.6>.

Pedersen, B. S., Nimb, S., Asmussen, J., Sørensen, N., Trap-Jensen, L., & Lorentzen, H. (2009). DanNet – the challenge of compiling a WordNet for Danish by reusing a monolingual dictionary. *Language Resources and Evaluation, Computational Linguistics Series*. 10.1007/s10579-009-9092-1. 31 s.

Pedersen, B. S., Sørensen, N. C. H., Olsen, S., & Nimb, S. (2024b). Evaluering af sprogforståelsen i danske sprogmodeller – med udgangspunkt i semantiske ordbøger. In *Nydanske Sprogstudier*, NyS 65, 8–40. <https://doi.org/10.7146/nys.v1i65.143072>

Pedersen, B. S., Sørensen, N. C. S., Olsen, S., Nimb, S., & Gray, S. (2024a). Towards a Danish Semantic Reasoning Benchmark - Compiled from Lexical-Semantic Resources for Assessing Selected Language Understanding Capabilities of Large Language Models. In N. Calzolari, M. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of LREC-COLING 2024*, Torino, Italy.

Sørensen, N. C. H., Nimb, S., & Pedersen, B. S. (2023). How do We Treat Systematic Polysemy in Wordnets and Similar Resources? – Using Human Intuition and Contextualized Embeddings as Guidance. In G. Rigau, F. Bond, & A. Rademaker (Eds.), *Proceedings of Global WordNet Conference 2023*, San Sebastián, Spain: Global WordNet Association (pp. 117–126).

Vossen, P. (Ed.). 1999. *EuroWordNet, A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers. <https://doi.org/10.1017/S1351324903223299>

Acknowledgements

The database structure of COR.SEM has been developed by Thomas Troelsgård, Det Danske Sprog- og Litteraturselskab.

Contact information

Sanni Nimb

Society for Danish Language and Literature
sn@dsl.dk

Ida Flörke

Society for Danish Language and Literature
if@dsl.dk

Sussi Olsen

Centre for Language Technology, NorS, University of Copenhagen
sol@hum.ku.dk

Bolette S. Pedersen

Centre for Language Technology, NorS, University of Copenhagen
bspedersen@hum.ku.dk

Nathalie C. H. Sørensen

Society for Danish Language and Literature
nats@dsl.dk

Andrej Perdih, Kozma Ahačič, Nataša Jakop, Nina Ledinek,
and Špela Petric Žižić

SEMANTIC INFORMATION ON THE FRANČEK EDUCATIONAL LANGUAGE PORTAL FOR SLOVENIAN

Abstract This article presents semantic information about contemporary standard Slovenian on the Franček educational language portal, which is aimed at primary- and secondary-school students. The portal's primary role is to enhance students' dictionary skills as part of the national language education program and to introduce users to other linguistic resources, such as school grammars. The portal offers a user-friendly environment tailored to meet the needs of its target users. Emphasis is placed on its innovative design and user-centric approach, which facilitate intuitive learning and engagement with lexicographic content. The lexicographic content of the Franček portal is organized in eight content modules altogether, and the semantic information discussed here is presented in meaning, synonymy, and phraseology modules. The article discusses the decision-making process, technical issues, and visualization strategies for adapting semantic information to three distinct age groups.

Keywords pedagogical lexicography; semantics; language portal; Slovenian language; dictionary linking

1. Introduction

Franček (www.francek.si) is a freely available educational language portal for Slovenian aimed at primary- and secondary-school students (Ahačič et al., 2021; Gliha Komac & Perdih, 2023; Perdih, 2022; Perdih et al., 2021). It is the first comprehensive interactive e-tool for Slovenian, and it has an original design not inspired by other language portals. It was developed to meet the basic objectives of the Slovenian primary-school and high-school curriculum because Slovenian linguists have repeatedly expressed the need for specialized high-quality educational language resources with reliable content, including grammars, dictionaries, e-tools, and portals (Godec Soršak, 2015; Rozman, 2010, 2012; Rozman et al., 2010; Stabej et al., 2008). The main goal of the portal is therefore to familiarize students with various types of dictionaries, grammars, and other language resources that they will later encounter as competent adult users. It aims to encourage young people to use dictionaries through intuitive practical use on a daily basis.

The portal's central component is a dictionary section with approximately 100,000 headwords. It is linked to the grammatical component (Ledinek, Ahačič, et al., 2022) and to the language consulting service for teachers. Teaching materials, a web application for automatically generating exercises (Godec Soršak et al., 2023), a corpus of Slovenian school texts (Ledinek, Jemec Tomazin, et al., 2022), and a mobile game

were also designed along with the portal. The portal was compiled in 2017–2021, but content updates and maintenance have been ongoing, albeit at a slower pace.

This article deals with semantic information on contemporary standard Slovenian on the Franček portal. It first presents the process of database compilation, then describes semantic information in individual modules, and also discusses the decision-making process with regard to adapting semantic information to target users.

2. Database Compilation

The dictionary section is organized not by individual dictionaries, but by content, based on the most common questions young people have about words. The presentation of dictionary data follows the principles of natural language learning. In each module, specific questions (e.g., *What does this word mean?*) commonly asked by language learners are explained with answers that meet the needs of the learning process and the cognitive and linguistic abilities of learners of different ages.¹ The portal gradually introduces learners to increasingly complex linguistic content.

One of the principal deficiencies of educational e-tools for Slovenian that were developed prior to the advent of the Franček portal, e.g., Pedagogical Grammar Portal (PSP) – was their limited scope. Consequently, users were unable to rely on them to contain the majority of the information they would expect, given the design of the e-tool. For the Franček portal, our objective was to provide users with answers to the majority of questions they would typically ask themselves.

Therefore, the headword list with 100,006 single-word headwords was constructed based on two general monolingual dictionaries: eSSKJ and SSKJ2. The lexicographic database was then constructed from preexisting monolingual dictionaries and databases for general (adult) users, as well as from new lexicographic data prepared specifically for creating the Franček portal. A total of thirteen dictionary databases were used.

Automatic linking was undertaken based on matching headwords' spelling, part-of-speech data, and stress.² Manual linking was required in other cases, especially in the case of homographs.

Dictionary databases were linked at the headword level, but no attempt was made to link them at the microstructural level, as in the case of the Estonian EKI Combined Dictionary (Tavast et al., 2020).³ The portal also differs from other lexicographic

¹ Adaptations are based on consultations and written and oral interviews with 72 primary and secondary schools teachers and their students, supplemented by insights from six experts in language didactics from the University of Ljubljana and the National Education Institute Slovenia.

² Historical dictionaries required thorough manual checking due to various changes caused by language development and orthographic changes.

³ Although linking at the sense level between the synonym dictionary (SSSJ) and the general monolingual dictionary (SSKJ2) has been ongoing (cf. Treven, 2022), its full realization surpassed the possibilities within the project of compiling the Franček portal. Linking other dictionaries at the microstructural level, however, would offer little benefit to the end users.

resources constructed from various dictionaries and other lexicographic databases—for example, DWDS (Klein & Geyken, 2010)—in that the content and visualization have been adapted for school students.

Because students differ significantly in terms of age, cognitive abilities, and needs, the content and visualization of the dictionary data were further adapted to three age groups:

1. first through fifth grade (ages six to ten),
2. sixth through ninth grade (ages eleven to fourteen) of primary school, and
3. high school (ages fifteen to eighteen).

The large majority of dictionaries used for compiling the database have been (or are being) compiled at the ZRC SAZU Fran Ramovš Institute of the Slovenian Language. This is why we were able to develop the portal without complex legal procedures. However, there are some exceptions, such as historical dictionaries for which copyright has expired or rights were obtained from the author's descendants, and part of the morphological database obtained from the IT company Amebis, which handled the technical part of the portal. All resources created specifically for Franček portal are available under open licences.

3. Semantic Information on the Portal

The lexicographic data on the portal are organized into eight content modules: meaning, synonymy, morphology, pronunciation, phraseology, dialect, etymology, and history (Figure 1). All modules are presented to the second and third age groups (ages eleven to eighteen), whereas presentation to users six to ten years old is limited to meaning, synonymy, morphology, and pronunciation modules.

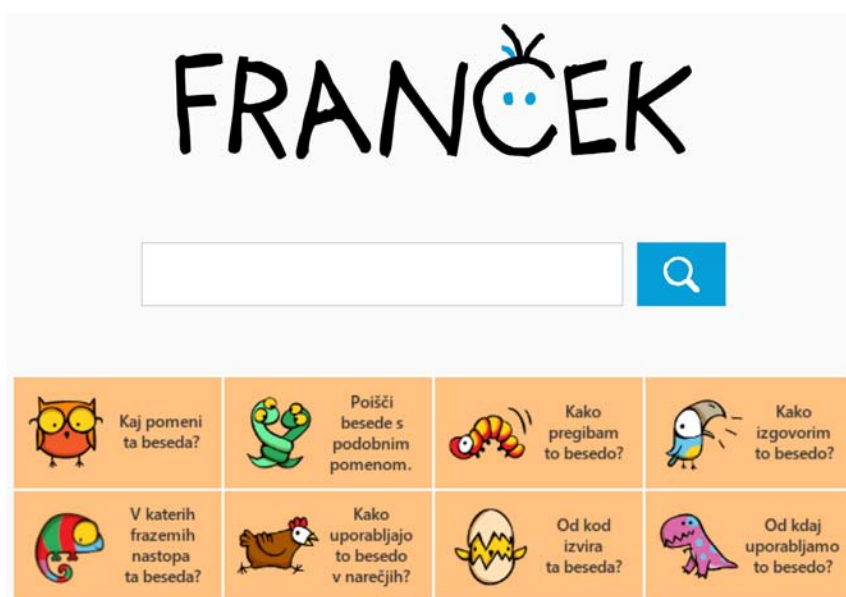


Fig. 1: The search field and eight lexicographic modules, second age group (<https://www.franček.si/>)

This article focuses on the semantics of contemporary standard language presented in the meaning, synonymy, and phraseology modules, and semantic indicators offered in the search process. Content in other modules, including the dialect module, which was presented at a previous Euralex conference (Ježovnik et al., 2020), surpasses the scope of the topic.

3.1 Meaning Module

The meaning module answers the question *What does this word mean?* and includes definitions and illustrative material, and it may also contain labels, images, and multiword units (Figure 2).

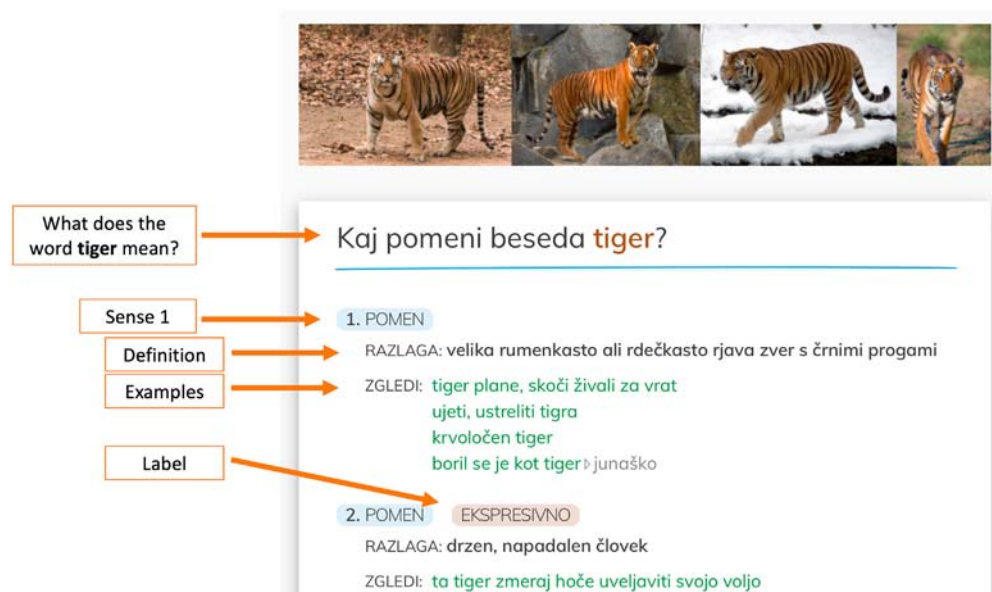


Fig. 2: The word *tiger* in the meaning module, second age group (<https://www.franček.si/beseda/s078836/tiger>)

The lexicographic data were obtained from three dictionaries. Their priority was agreed as follows:

1. ŠSSJ for the first age group (ages six to ten) only,
2. eSSKJ, and
3. SSKJ2.

When an entry was not available in a dictionary with higher priority, data from a dictionary with lower priority were used. In this way, the youngest users were primarily given data from the new School dictionary of the Slovenian language (ŠSSJ; Godec Soršak, 2022; Petric Žižić, 2020, 2022), currently containing 3,087 entries. If an entry did not exist in this dictionary, data from eSSKJ were used, which is also the primary dictionary source for the second and third age groups (ages eleven to fourteen, and fifteen to eighteen). Because eSSKJ and the School Dictionary of the Slovenian language are still being compiled, the data from (the only completed) dictionary of standard Slovenian (SSKJ2) are provided when entries are not yet available in these dictionaries.

3.1.1 School Dictionary

The main purpose of ŠSSJ is pedagogical; its primary aim is to familiarize students with the structure of a dictionary, and it has therefore been designed taking into account the recommendations of teaching experts and surveys completed by fifth-grade students (ten years old) in class.⁴ The source of materials for ŠSSJ is the Corpus of Slovenian school texts, which encompasses texts from textbooks, children's literature by select authors, and school essays. The lexis examined in the dictionary is thus taken from texts that are the main source of new standard vocabulary for students at this age (Petric Žižić, 2022, p. 128).

ŠSSJ provides sample terminology from fundamental subject fields that students are expected to learn by age ten. General lexis is included according to three principles. The basic, more frequently used lexemes familiarize the user with the structure of the dictionary and with the properties of the lexeme in general (e.g., pronunciation, inflection, meaning, and polysemy). The second principle of inclusion is based on the criteria of problem selection, which includes lexemes that are potentially difficult to inflect or spell. The third group of headwords contains lexemes with abstract meanings (e.g., *način* 'manner', *lastnost* 'property', and *logika* 'logic'). The dictionary also includes multiword lexical units with pedagogically adapted definitions.

All types of dictionary definitions in ŠSSJ come in the form of sentences.⁵ In line with the established principles of Slovenian lexicography, the explanatory part of the sentence in autosemantic headwords is based on *genus proximum* and *differentia specifica*; for example, *notebook*: a **notebook** is a booklet with pages and binding used to write or draw in during class. The conditional-clause type of definition is used for explaining verbal set phrases and phraseological units; for example, *to keep company*: If someone **keeps company** with someone, it means they hang out and spend time together.⁶

ŠSSJ describes only the meanings that appear in the corpus; it therefore only includes meanings that are relevant to students. When defining a meaning, essential semantic features are complemented with additional circumstances of the denotatum's occurrence, as would normally be the case. The language of the definitions is less complex (especially at the lexical and syntactic levels), taking into account the principles of naturalness and comprehensibility (in terms of mimicking spontaneous speech, particularly in the syntactic structure of a definition, sometimes

⁴ The first type of question related to the structure of the dictionary and the basic vocabulary (e.g., *How many meanings does the word have? Where are the examples of the use of the word in the dictionary? Which multiword lexical units does the word appear in?*). The second type of question was used, for example, to determine whether the terminological labels were sufficiently understandable and meaningful (e.g., *Which school subject does the second meaning of the word point occur in?*), which type of definition is the most understandable (students deduced which word the definition referred to and formed their own definitions of the given words), whether certain implicitly given information (e.g., the distinction between possessive and ordinal adjectives), interferes with understanding the definition, and so on. The predominance of correct answers confirmed that the conceptual solutions were appropriate.

⁵ Full-sentence definitions were introduced for the first time in Collins COBUILD English Dictionary (1987).

⁶ The definitions are translated.

at the expense of its economy). Definitions emphasize the most typical aspect of the meaning from a student's point of view; the use of the first-person plural serves as a means to make the content of the definitions more accessible through the user's experiences (Petric Žižić, 2020, p. 393).

3.1.2 eSSKJ and SSKJ2

The general monolingual dictionaries eSSKJ and SSKJ2 are aimed at adult users and provide comprehensive microstructural information (Gliha Komac et al., 2015; Perdih & Snoj, 2015). Consequently, the data from the dictionaries were selected and abridged to better suit the target users. This solution was accepted even for the third age group because the volume of such data can be very extensive. The decision was made because the primary purpose of Franček is to facilitate young users' transition to using available unadapted dictionaries, and therefore a link is always provided to the original dictionary available on the Fran dictionary portal (Ahačič et al., 2015; Perdih, 2020).

The eSSKJ dictionary was partially adapted to achieve greater consistency with the data from SSKJ2, most notably by reducing the number of examples (collocations and example sentences) and omitting detailed information on syntactic structures.

In SSKJ2, minor semantic shifts and rare syntactic structures were omitted, and the amount of illustrative material was reduced similarly as in the case of eSSKJ. For the first two age groups, the scope of terminology was adjusted to the school curriculum, thus further reducing the amount of data presented. Labels that were abbreviated in SSKJ2 are written out in full.

Because eSSKJ and SSKJ2 (as well as Dictionary of Slovenian Phrasemes (SSF) in phraseological module) interpret labels differently, we conceptually harmonized them and presented their unified descriptions in the form of tooltips with links to the school grammar, where further information is provided.

Labels pertaining to all or to the majority of senses in SSKJ2 are presented in the head of the dictionary entry; that is, at the entry level. Because presentation of data from the head is limited, and to offer more understandable information to the end-user in the semantic module itself, entry-level labels were transferred to the sense level. Special treatment was necessary in the case of various combinations between the head-level label and sense-level label.

Designations of microstructural elements, such as definitions and illustrative material, are provided for the second age group (Figure 2) to help users obtain a clear idea about the main components of the dictionary structure. The third age group, on the other hand, may no longer need this help, therefore such information is omitted or limited, making the visualization closer to dictionaries for adult users. Similar approach can be observed in other modules (Gliha Komac & Perdih, 2023, p. 18).

3.1.3 Images

To better support understanding of meanings, 64,063 images are provided for 19,760 entries. The majority of images are license-free images obtained via the Bing search engine, and the Slovenian Ethnographic Museum gave permission to use photos from its gallery. All images were subsequently manually selected by a group of students for inclusion in the database. In cases of polysemy, an effort was made to find images for several senses. At the time of image selection activities, AI tools for image generation were not yet available.

3.2 Synonymy Module

The synonymy module provides answers to the question *Which words have similar meanings to the word X?* and contains definitions, synonyms, and icons and/or a textual description representing labels.

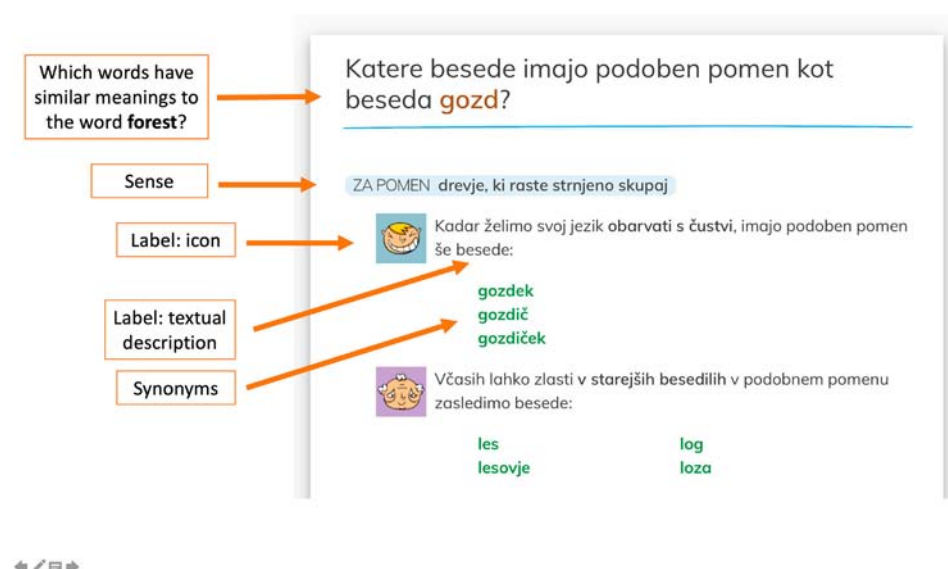


Fig. 3: The word *gozd* 'forest' in the synonymy module, second age group (<https://www.franček.si/beseda/s018093/gozd>)

To enhance students' awareness that a single meaning in standard Slovenian can be expressed with different words, taking into account the different circumstances of communication, the Franček portal introduces students to a number of synonyms of modern standard language. The source for this is the Synonym Dictionary of Slovenian Language (SSSJ). The synonymy module utilizes stylistic and register labels from the source dictionary to illustrate how the synonyms identified differ from each other and when it is appropriate to use them in different texts and language situations.

The information from the Slovenian synonym dictionary has been adapted for the educational portal in terms of both content and presentation. Lexemes are introduced with a dictionary definition, after which synonyms are assigned to

them. Subsequently, the synonyms are grouped according to the markedness of the synonym. For neutral unmarked lexemes such as *gozd* 'forest', all eligible synonyms are listed. For marked synonyms, the synonym module redirects the user to an unmarked synonym.

The source dictionary describes synonyms with a relatively detailed set of labels. For the Franček target users, some types of synonyms were omitted based on their labels, and the remaining labels were merged into broader groups. To reinforce the students' understanding of standard language, the following types of synonyms are given:

1. neutral synonyms (without stylistic markedness), which are not restricted in distribution and are used in everyday practical communication,
2. synonyms with a positive expressive value (i.e., endearing, childish, humorous, and elated expressions),
3. synonyms with a negative expressive value (i.e., derogatory and pejorative expressions), and
4. time- and frequency-marked synonyms (i.e., archaic, obsolete, and rare expressions).

The Franček educational portal primarily familiarizes users with the characteristics of the standard language. Because they are not part of the standard language, colloquial and jargon synonyms of the lexeme searched for are not shown in the synonym module. Similarly, the module does not include dialect synonyms (these are shown in a separate module) or terminological synonyms (terminology is a specific diasystem of language, and therefore the relationship between terminology and general language is not a matter of stylistic choice). Because the portal is intended for the school-age population, vulgar synonyms have also been excluded.

The presentation of data in the synonym module is adapted to students' cognitive abilities and varies for each age level. Markedness of synonyms from the above mentioned four groups is presented as icons and/or textual description. The visualization for students in the first age group includes only icons alongside the synonyms, which intuitively convey whether and in what circumstances the use of a synonym is acceptable (e.g., a thumbs-up on a green background indicates unmarked synonyms, and a frowning face on a red background indicates synonyms with negative expressive value). The visualization for the second age group includes both icons and typified descriptions with information about the stylistic value of the lexemes. The visualization for the third age group contains only a textual description.

3.3 Phraseology Module

The phraseology module provides answers to the question *Which phrasemes does the word appear in?* and contain definitions, illustrative material, and variations

in phraseological units. The primary aim of the phraseological module is didactic and informative. Therefore, adaptations required consideration of both macrostructural and microstructural aspects: which phraseological units are to be included, in which dictionary entries they are found, and how they are to be presented.



Fig. 4: The word *buča* 'pumpkin' in the phraseology module, second age group (<https://www.franček.si/beseda/e000495/buča>)

The phraseology module is available only for the second and third age groups. The first age group also encounters phraseological units, but only within the scope of the ŠSSJ in the meaning module. For this age group, we introduce students to some basic phraseological features, such as multiword fixed structures and figurative meanings. For the second age group (as shown on Figure 4), we present phraseology as a standalone module in more detail, including a higher number of phraseological units recognized as basic forms in source dictionaries, and their variants. For the third age group, we expand the set of variants to offer users a comprehensive picture of each phraseological unit.

The data for the phraseology module come from the SSF and eSSKJ. Various types of phraseological units are included and presented: nominal, adjectival, verbal, adverbial, and sentence phraseological units, including paremiological units. They are arranged alphabetically by the first component, with non-sentential phraseological units presented first, followed by sentence phraseological units, which, unlike the others, are formally recognizable due to notation with a capital letter. Each phraseological unit is shown in its basic dictionary form (phraseological lemma), accompanied by a semantic description, and variants of the phraseological units are also listed. This ensures that students receive the key information about the use of each phraseological unit and its variants.

3.4 Semantic Indicators

To help distinguish between homographs in the search process, semantic, part-of-speech, and other types of indicators⁷ are provided (Perdih, 2021). Part-of-speech indicators are given to 1,165 headwords, but semantic indicators are predominantly used (3,268), and in the case of fifty-seven headwords other solutions were required. To differentiate between homographs pertaining to the same part of speech, semantic indicators are used, typically in the form of synonyms or hypernyms, and in some cases the specification of domain (e.g., *kubičen: v matematiki* ‘cubic: in mathematics’) or a prepositional phrase⁸ is used (e.g., *akorden: o plačilu* ‘jobbing: about payment’).

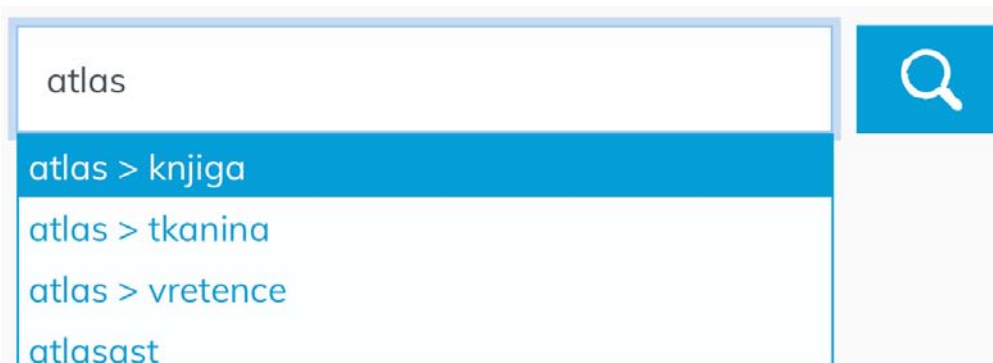


Fig. 5: The homonyms *atlas* with semantic indicators (‘book’, ‘fabric’, ‘vertebra’) in the search field (<https://www.franček.si/>)

4. Conclusion

The semantic information on contemporary standard Slovenian constitutes a central point of lexicographic description on the Franček educational language portal. The meaning and synonymy modules, as well as the phraseology module, provide users with definitions for single- and multi-word expressions, synonyms, examples, labels, links to grammar and other features. The data were sourced from existing explanatory, synonymy, or phraseology dictionaries and from a school dictionary that was specifically compiled for this project. The goal was to adapt the content and visualization to better suit the needs of the target users.

The primary objectives of the Franček educational portal are to provide students with language resources that would facilitate the initial steps in using dictionaries, assist students in comprehending the content of dictionaries, inform about fundamental language features, and to raise awareness of dictionaries as reliable companions for resolving language issues. The modifications of content and visualization were made gradually for three age groups with the intention of facilitating a more seamless transition from school dictionaries to dictionaries for adults.

⁷ In contrast to the common use of indicators, especially semantic ones, that are used for sense distinction and can be used to build sense menus of a single entry (cf. Kosem et al., 2013; Lew & Pajkowska, 2007), here the distinction between homographs is the main goal.

⁸ Cf. Gantar (2015, pp. 168–169) for implementation of this type of indicators in the Slovene lexical database.

References

- Ahačič, K., Ježovnik, J., Ledinek, N., Perdih, A., Petric, Š., & Race, D. (2021). Priprava jezikovnih podatkov za pedagoški portal o slovenščini Franček. *Philological Studies*, 19(1), 203–224. <https://doi.org/10.17072/1857-6060-2021-19-1-203-223>
- Ahačič, K., Ledinek, N., & Perdih, A. (2015). Fran: Next Generation Slovenian Dictionary Portal. In K. Gajdošová & A. Žáková (Eds.), *Natural language processing, corpus linguistics, lexicography / Eighth International Conference, Bratislava, Slovakia, 21-22 October 2015: Proceedings* (pp. 9–16). RAM-Verlag.
- COBUILD 1: *Collins COBUILD English Dictionary*, Edition 1 (1987). HarperCollins.
- eSSKJ: *Slovar slovenskega knjižnega jezika*. (2016–). ZRC SAZU, Založba ZRC. Retrieved May 29, 2024, from <https://www.fran.si>
- Franček. Retrieved May 29, 2024, from <https://www.franček.si/>
- Gantar, P. (2015). *Leksikografski opis slovenščine v digitalnem okolju*. Znanstvena založba Filozofske fakultete Univerze v Ljubljani. <https://doi.org/10.4312/9789612377922>
- Gliha Komac, N., & Perdih, A. (2023). Slovenian language resources at your fingertips: Fran, Franček, Terminologično, and Language advisory. In G. Haja & M. Nastasia (Eds.), *Actele Colocviului Internațional Lexicografia academică românească. Provocările informatizării 2023* (pp. 11–26). Universitatea Babeș-Bolyai, Presa Universitară Clujeană. <https://doi.org/10.52257/30086159>
- Gliha Komac, N., Jakop, N., Ježovnik, J., Klemenčič, S., Krvina, D., Ledinek, N., Mirtič, T., Perdih, A., Petric, Š., Snoj, M., & Žele, A. (2015). *Koncept novega razlagalnega slovarja slovenskega knjižnega jezika*. Inštitut za slovenski jezik Frana Ramovša ZRC SAZU. https://fran.si/179/novi-slovar-slovenskega-knjiznega-jezika/datoteke/Potrjeni_koncept_NoviSSKJ.pdf
- Godec Soršak, L. (2015). Slovenski otroški šolski slovar. In M. Smolej (Ed.), *Slovnica in slovar—Aktualni jezikovni opis* (Vol. 1, pp. 243–250). Znanstvena založba Filozofske fakultete. https://centerslo.si/wp-content/uploads/2015/11/34_1-Godec-Sor.pdf
- Godec Soršak, L. (2022). Teoretični pogled na šolski slovar. *Slavistična Revija*, 70(4), 667–679. <https://doi.org/10.57589/srl.v70i4.4090>
- Godec Soršak, L., Ledinek, N., Bizjak, N., & Rot Vrhovec, A. (2023). Didaktično gradivo na portalu Franček. *Jezikoslovni Zapiski*, 29(2), 225–247. <https://doi.org/10.3986/JZ.29.2.09>
- Ježovnik, J., Kenda-Jež, K., & Škofic, J. (2020). Reduce, Reuse, Recycle: Adaptation of Scientific Dialect Data for Use in a Language Portal for Schoolchildren. In Z. Gavrilidou, M. Mitsiaki, & A. Fliatouras (Eds.), *Lexicography for Inclusion: Proceedings of the 19th EURALEX International Congress, 7-9 September 2021, Alexandroupolis, Vol. 1* (pp. 31–37). Democritus University of Thrace. https://www.euralex.org/elx_proceedings/Euralex2020-2021/EURALEX2020-2021_Vol1-p031-037.pdf
- Klein, W., & Geyken, A. (2010). Das Digitale Wörterbuch der Deutschen Sprache (DWDS). *Lexicographica*, 26(2010), 79–96. <https://doi.org/10.1515/9783110223231.1.79>

Kosem, I., Gantar, P., & Krek, S. (2013). Automation of lexicographic work: An opportunity for both lexicographers and crowd-sourcing. In I. Kosem, J. Kallas, P. Gantar, S. Krek, M. Langemets, & M. Tuulik (Eds.), *Electronic lexicography in the 21st century: Thinking outside the paper. Proceedings of the eLex 2013 conference, 17-19 October 2013, Tallinn, Estonia* (pp. 32–48). Trojina, Institute for Applied Slovene Studies; Eesti Keele Instituut.

Ledinek, N., Ahačič, K., & Petric, Š. (2022). Presenting and linking grammatical data on the Franček educational language portal. *Slovenski Jezik - Slovene Linguistic Studies*, 14, 155–182. <https://doi.org/10.3986/sjsls.14.1.06>

Ledinek, N., Jemec Tomazin, M., Trojar, M., Perdih, A., Ježovnik, J., Romih, M., & Erjavec, T. (2022). Korpus šolskih besedil slovenskega jezika: Zasnova in gradnja. *Jezikoslovni Zapiski*, 28(1), 122–137. <https://doi.org/10.3986/JZ.28.1.07>

Lew, R., & Pajkowska, J. (2007). The effect of signposts on access speed and lookup task success in long and short entries. *Horizontes de Lingüística Aplicada*, 6(2), 235–232.

Perdih, A. (2020). Portal Fran: Od začetkov do danes. *Rasprave Instituta Za Hrvatski Jezik i Jezikoslovlje*, 46(2), 997–1018. <https://doi.org/10.31724/rihjj.46.2.28>

Perdih, A. (2021). Indikatorji pri homografih na portalu Franček. *Jezikoslovni Zapiski*, 27(2), 7–21. <https://doi.org/10.3986/JZ.27.2.01>

Perdih, A. (2022). Pedagoški leksikografski portal Franček. In M. I. Chernysheva (Ed.), *Sovremennoe razvitie slavjanskoj leksikologii i leksikografii* (pp. 340–354). Institut ruskogo jazyka im. V. V. Vinogradova RAN. https://doi.org/10.31912/slavic_lexicology.2022.340-354

Perdih, A., & Snoj, M. (2015). SSKJ². *Slavia Centralis*, VIII(1), 5–15.

Perdih, A., Ahačič, K., Ježovnik, J., & Race, D. (2021). Building an Educational Language Portal Using Existing Dictionary Data. *Journal of Linguistics/Jazykovedný Časopis*, 72(2), 568–578. <https://doi.org/10.2478/jazcas-2021-0052>

Petric Žižić, Š. (2020). Tipologija razlag v Šolskem slovarju slovenskega jezika. *Slavistična Revija*, 68(3), 391–409.

Petric Žižić, Š. (2022). School dictionary of the Slovenian language on the Franček web portal. *Slavistica Vilnensis*, 67(2), 126–140. [https://doi.org/10.15388/SlavViln.2022.67\(2\).100](https://doi.org/10.15388/SlavViln.2022.67(2).100)

PSP: *Pedagoški slovnični portal*. Retrieved May 29, 2024, from <https://slovnica.slovenscina.eu/>

Rozman, T. (2010). *Vloga enojezičnega razlagalnega slovarja slovenščine pri razvoju jezikovne zmožnosti: Doktorska disertacija*. Univerza v Ljubljani.

Rozman, T. (2012). Jezikovni pouk slovenščine: Model (za) nove generacije. In B. Krakar-Vogel (Ed.), *Slavistika v regijah—Koper* (pp. 219–224). Zveza društev Slavistično društvo Slovenije; Znanstvena založba Filozofske Fakultete.

Rozman, T., Stritar Kučuk, M., Krapš Vodopivec, I., Kosem, I., & Krek, S. (2010). *Nova didaktika poučevanja slovenskega jezika: Sporazumevanje v slovenskem jeziku* (p. 664). Ministrstvo za

šolstvo in šport; Amebis. http://projekt.slovenscina.eu/Media/Kazalniki/Kazalnik15/Nova_didaktika_Sporazumevanje.pdf

SSF: Keber, J. *Slovar slovenskih frazemov*. (2015). Založba ZRC, Znanstvenoraziskovalni center SAZU. Retrieved May 29, 2024, <https://www.fran.si>

SSKJ2: *Slovar slovenskega knjižnega jezika*. (2014). Druga, dopolnjena in deloma prenovljena izdaja. Založba ZRC, Znanstvenoraziskovalni center SAZU. Retrieved May 29, 2024, from <https://www.fran.si>

ŠSSJ: Petric Žižić, Š. (Ed.). *Šolski slovar slovenskega jezika*. Franček. (2021–). ZRC SAZU. Retrieved May 29, 2024, from <https://www.francek.si>

SSSJ: *Sinonimni slovar slovenskega jezika*. (2018). ZRC SAZU, Založba ZRC. Retrieved May 29, 2024, from www.fran.si

Stabej, M., Rozman, T., Pirih Svetina, N., Modrijan, N., & Bajec, B. (2008). *Jezikovni viri pri jezikovnem pouku v osnovni in srednji šoli: Končno poročilo z rezultati dela*. Pedagoški inštitut. <http://www.trojina.si/wp-content/uploads/2019/08/StabejRozman.pdf>

Tavast, A., Koppel, K., Langemets, M., & Kallas, J. (2020). Towards the Superdictionary: Layers, Tools and Unidirectional Meaning Relations. In Z. Gavrilidou, M. Mitsiaki, & A. Fliatouras (Eds.), *Lexicography for Inclusion: Proceedings of the 19th EURALEX International Congress, 7-9 September 2021, Alexandroupolis, Vol. 1* (pp. 215–223). Democritus University of Thrace. https://www.euralex.org/elx_proceedings/Euralex2020-2021/EURALEX2020-2021_Vol1-p215-223.pdf

Treven, M. (2022). *Medslovarsko pomensko povezovanje na primeru SSSJ in SSKJ 2: Nabor iztočnic s sklicevalno razlago glagolnik od X: diplomsko delo*. Univerza v Ljubljani, Filozofska fakulteta.

Acknowledgements

This article was written within the Slovenian Research Agency's P6-0038 The Slovenian Language in Synchronic and Diachronic Development program group.

Contact Information

Andrej Perdih

ZRC SAZU, Fran Ramovš Institute of the Slovenian Language
andrej.perdih@zrc-sazu.si

Kozma Ahacič

ZRC SAZU, Fran Ramovš Institute of the Slovenian Language
kozma.ahacic@zrc-sazu.si

Nataša Jakop

ZRC SAZU, Fran Ramovš Institute of the Slovenian Language
natasa.jakop@zrc-sazu.si

Nina Ledinek

ZRC SAZU, Fran Ramovš Institute of the Slovenian Language
nina.ledinek@zrc-sazu.si

Špela Petric Žižić

ZRC SAZU, Fran Ramovš Institute of the Slovenian Language
spela.petric@zrc-sazu.si

Emma Sköldberg, Shafqat Mumtaz Virk, Pauline Sander,
Simon Hengchen, and Dominik Schlechtweg

REVEALING SEMANTIC VARIATION IN SWEDISH USING COMPUTATIONAL MODELS OF SEMANTIC PROXIMITY

Results From Lexicographical Experiments

Abstract The paper reports a pilot study on the detection of lexical semantic variation in modern Swedish. The starting point of the study is the meaning descriptions of around 65,000 headwords in 'The Contemporary Dictionary of the Swedish Academy' (SO, 2021) covering approximately 100,000 different senses. In our work, we aim to explore the potential of the latest computational methods to discover outdated definitions in SO and update them. For this, we make use of the DUREl tool (Schlechtweg et al., 2018, 2024) which relies on state-of-the-art language models for the automatic semantic analysis of word usages. The work resulted in drawing lexicographers' attention to both main senses and subsenses that should be added to the dictionary. It has also demonstrated that certain meaning descriptions in SO are too general and should be split in accordance with the current principles for the semantic descriptions in the dictionary.

Keywords lexicography; language technology; lexical semantics; semantic variation and change; automatic semantic analysis; word sense induction, semantic proximity; Swedish; DUREl

1. Introduction

As shown by e.g., Lau et al. (2012), the meaning descriptions in dictionaries frequently become outdated, and hence require continuous updates. In order to meet the needs in practical lexicography in relation to this area, Cook et al. (2013), Nimb et al. (2020) among others, have conducted lexicographically-oriented studies on the automatic detection of semantic variation and change. However, the research in this direction, including method development could be taken much further, especially in relation to relatively under-resourced languages such as Swedish (cf. Cavallin, 2012).

In this paper, we report on ongoing research regarding the automatic detection of lexical semantic variation and change in modern Swedish, and how this work benefits Swedish lexicography. The starting point of the experiments in the study is the meaning descriptions in *Svensk ordbok utgiven av Svenska Akademien* ('The Contemporary Dictionary of the Swedish Academy' (SO, 2021) (see e.g., <https://svenska.se/>), which constitutes the most complete description of the Swedish vocabulary of today (see Section 2).

In this work, we aim to explore the potential of the latest computational methods to discover outdated definitions in SO and update them. For this, we make use of the annotation tool *DURel* (Schlechtweg et al., 2018, 2024) which relies on state-of-the-art language models for automatic semantic analysis of word usages (see Section 3). In more concrete terms, we have conducted two experiments. The first experiment aimed to validate the feasibility and prove the usefulness of *DURel* and tune hyper-parameters of its underlying computational language models. The second experiment aimed, by using *DURel*, to revise semantic descriptions of headwords already included in SO.

The outline of the paper is as follows: In Section 2 we present SO with focus on the meaning descriptions in the dictionary. Section 3 includes a brief overview of the annotation tool *DURel* as well as the corpus that we have been using in the experiments. After that, in Section 4, we present the two experiments that we have conducted and finally, in Section 5, we summarize and discuss our results.

2. SO and Its Sense Descriptions

As already mentioned, the point of departure in our research is SO, a definition dictionary with about 65,000 headwords describing the vocabulary of modern Swedish. SO, which is corpus-based, covers approximately 100,000 different senses and it is primarily aimed at users with Swedish as their mother tongue, but also advanced language learners of Swedish.

SO is a subset of a very extensive lexical database, which has been under continual development at the University of Gothenburg since the 1970s (see Ralph et al., 1977). For each edition, SO will provide as complete information as possible on recurrent words and expressions in Swedish general language. This information includes the words' spelling, pronunciation, inflection, etymology etc. However, the emphasis in SO is on the meanings and uses of the words (see e.g., Malmgren & Sköldberg, 2013).

As is well-known, practical lexicographic work is tradition-bound and all the editions of SO follow certain principles described in e.g., Ralph et al. (1977) and Järborg (1989). These principles also apply to the meaning descriptions in the dictionary. According to Svensén (2009, pp. 211–212), the polysemy structure of the words in a dictionary can be described 1) linearly, i.e., as a number of discrete units arranged in a sequence or 2) hierarchically, as a number of main/core senses, to which groups of subsenses/shades are associated. In SO, the latter approach is adopted. Furthermore, in SO, the relationship between every main sense and the subsense(s) is explicitly specified (in terms of meaning extension, meaning specialisation, metaphorical (figurative) use etc.) To illustrate how the two different polysemy structure principles work, consider the example *aptitretare* (appetizer) with one more literal and one more figurative sense. When the polysemy structure is described linearly in the dictionary, these two meanings are listed as 'meaning 1' and 'meaning 2' and when the polysemy structure is hierarchical, the two meanings are listed as 'meaning 1' and 'meaning 1a', because the second meaning is considered to be a figurative subsense of the main sense.

As Lew (2013, p. 287), among others, have shown, two opposing strategies, known as *lumping* and *splitting* have been identified when it comes to specifying senses in monolingual dictionaries. The first strategy aims to minimize the number of senses so that each of them cover as much semantic ground as possible. The second strategy tends to result in a rather larger number of finely distinguished senses. The lexicographers' choice of strategy often depends on the current type of dictionary and the dictionary's target users (Atkins & Rundell, 2008, pp. 267–268). The semantic description in a dictionary that primarily addresses learners at beginner level is frequently characterized by lumping while the semantic description in a dictionary that addresses linguists, etc. is characterized by splitting. In SO, the analysis and the distinction into different senses is relatively “fine-grained”. In other words, the SO lexicographers are rather splitters than lumpers.

Since the semantics of the headwords in SO is fundamental, an important part of the revision work between the editions is to examine if the meanings of the headwords have changed – and in that case, how they have changed. However, this work is very time-consuming. As pointed out by Petersson & Sköldbberg (2021), the SO lexicographers currently do not use any computational methods for discovering semantic change in a systematic way.

3. Tool and Corpora

In this section, we shortly present the tool and the corpus that we have used in the experiments.

The DUREl annotation tool (<https://durel.ims.uni-stuttgart.de/>) provides functionalities to upload sets of word usages for a particular word of interest and to annotate the semantic proximity (a.k.a. similarity/relatedness) between these usages either with humans or computers. This information (usages with their proximities) can then be visualized in a graph and clustered into sets of high-proximity usages, which can be interpreted as word senses.

The tool allows to inspect the graph and to adjust multiple annotation, cluster and visualization parameters. In this way, we can easily infer lexicographically relevant semantic information such as the inferred number of senses or their changes over time. Find an example of a DUREl graph concerning the usage of the noun *baksida* ('back, downside, disadvantage') in the SVT corpus in Figure 1.

DUREl offers multiple computational annotators for semantic proximity based on state-of-the-art pre-trained Word-in-Context models (Pilehvar & Camacho-Collados, 2019). We rely on the most recent model, XL-LEXEME, which was optimized for Lexical Semantic Change Detection (Cassotti et al., 2023). This model maps each word usage onto a contextualized distributional-semantic vector representation and then uses cosine similarity to estimate the semantic proximity between two usage vectors.

For our experiments, we have used the SVT corpus which consists of texts published by the Swedish public service television company between 2004 and 2021. The corpus covers about 200 million tokens and is available through Korp, Språkbanken's word research platform (Borin et al., 2012).

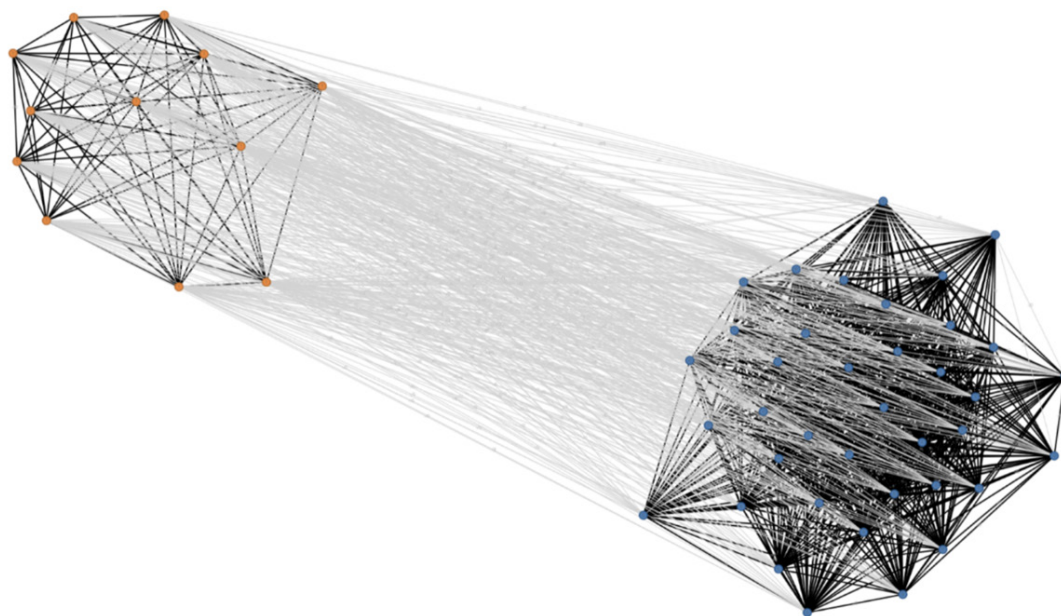


Fig. 1: The noun *baksida* ('back, downside, disadvantage'), based on corpus samples from the SVT corpus, forming two semantic clusters in DUREl.

Figure 1 visualizes usages of the word *baksida* as nodes while edge weights are given by the cosine similarity between usages annotated with the XL-LEXEME model. The graph was clustered using Correlation Clustering (Schlechtweg et al., 2020). DUREl highlights two semantic clusters, one with blue dots and another with orange dots among the uses of the word *baksida* in the corpus. If the users click on individual dots in the DUREl front-end interface, they will find the corresponding corpus sample. In this case, among the blue dots there are samples including word combinations such as “polishusets baksida” (‘the back of the police house’) and “baksidan av låret” (‘the back of the thigh’). Among the orange dots, there are samples such as “framgångens baksida” (‘lit. the back side of the success’) and “baksidan med droger” (lit. ‘the back side with drugs’). In other words, DUREl has here succeeded in pointing out at least two different senses of the word in the corpus, the latter of which is clearly metaphorical.

4. The Experiments

Petersson & Sköldberg (2021) show, by discussing different types of examples in Swedish, that the phenomenon of lexical semantic variation and change, is multifaceted and diverse. Hence, within our research, we strive to approach the phenomenon from partially different angles. One way is to start from SO dictionary articles concerning headwords with more than one sense and evaluate whether DUREl manages to cluster these senses among the uses of the word in a corpus (see experiment 1). Another way

is to start with headwords in SO with only one sense and see if DUREl predicts two or more sense clusters of the word in a corpus. Two or more clusters may indicate that the word has more than one sense in the texts and that the semantic description in SO should be revised (see experiment 2).

4.1 Experiment 1

In experiment 1, we started by selecting a set of 25 SO-headwords, based on their semantic characteristics: all of them have a main sense and one or more subsenses. An example is the adjective *enkelspårig* with the main sense ‘one-track’ and the figurative subsense ‘simplistic, superficial, narrow-minded’.

Next, 50 random usages (sentences) of each headword were extracted from the SVT corpus. After that, SO lexicographers assessed all the sentences and classified if the meaning of a certain word usage was covered in the dictionary and, if that was the case, which sense it was (the main sense, a subsense etc.). The goal of this work was partly to get an indication of the semantic quality of the current dictionary articles, partly to create a so-called manually curated gold standard (see below). This closer examination of the corpus samples showed, among other things, that the semantic description of 7 of the selected headwords in the dictionary were not comprehensive enough and should be supplemented. For example, in 12 out of 50 occurrences in the corpus, the verb *explodera* (‘explode’) meant ‘increase significantly’ (see e.g., ”Försäljningen av miljöbilar har exploderat det senaste året” (‘Sales of green cars have exploded in the past year’)). In the sample you find a well-established figurative use of the word *explodera* that is currently missing in SO and according to the editorial principles of the dictionary (see Ralph et al., 1977, and Järborg, 1989), that sense should form a subsense of its own in the dictionary article. Consequently, *explodera* and 6 other headwords were, based on quality reasons, excluded from the experiment. Even if the quality of SO is high in general, this result supports previous conclusions by e.g., Blensenius et al. (2021) concerning the development potential of the semantics in the dictionary.

After that, the remaining 18 headwords were uploaded to DUREl. Each of the 50 usages of the words was paired with every other usage producing a total of 2450 usage pairs for each headword. These pairs were then annotated automatically for sense similarity (i.e., if a given headword is used in the same sense in the usage pair or not) using XL-LEXEME (Cassotti et al., 2023), a state-of-the-art language model for lexical semantic change detection. The annotations were clustered with the correlation clustering algorithm described in Schlechtweg et al. (2020). The clustering algorithm has a threshold hyper-parameter that was tuned to yield optimal results on the gold data (see below). After uploading the usages, all further steps were performed using the DUREl front-end interface. Figure 2 shows the two resulting clusters for *enkelspårig*, one with blue dots and another with orange dots, corresponding to the literal and figurative senses of this headword in SO.

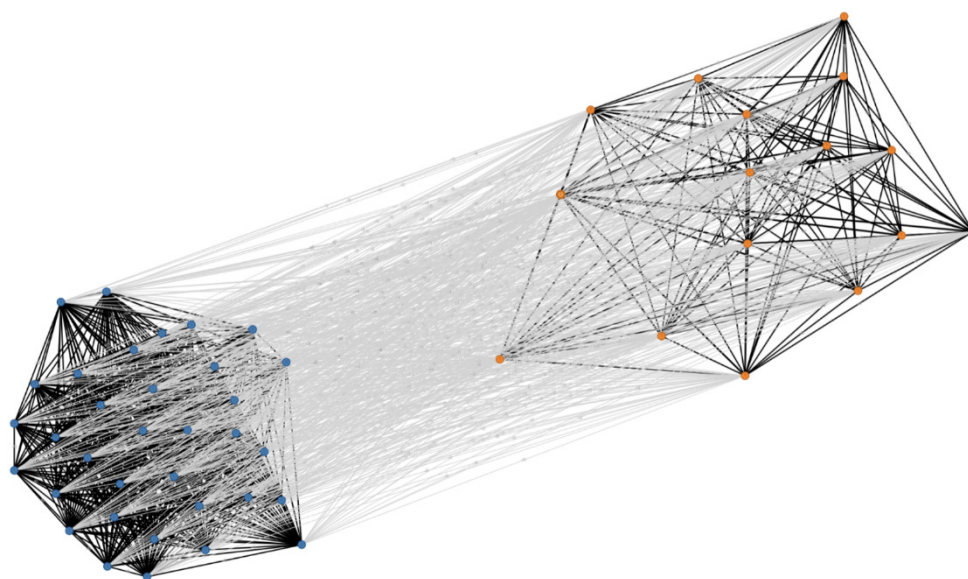


Fig. 2: DUREl sense cluster for the headword *enkelspårig* (‘one-track; ‘simplistic, superficial, narrow-minded’).

In the front-end interface, you see, by clicking on the blue spots, corpus samples like “Den enkelspåriga järnvägen mellan Motala och Hallsberg är idag en flaskhals ...” (‘The one-track railway between Motala and Hallsberg is a bottleneck today ...’). Hence, in this particular context, the literal, main sense of the adjective is used. By clicking on orange dots, you see uses like “De tror att vi är enkelspåriga lantisar, de tror att vi är trångsynta, att vi är rasister och homofober.” (‘They think we’re narrow-minded peasants, they think we’re bigoted, that we’re racists and homophobes.’). Here, the figurative sense of the word, which corresponds to the subsense in SO, is relevant. To conclude, in this case, DUREl has managed to distinguish between the two senses of *enkelspårig*.

In the next step of the experiment, the gold clusters, i.e., the results of the manual analysis of the language samples by the lexicographers, were evaluated against the automatically generated clusters by DUREl. Table 1 shows the Adjusted Rand Index (ARI; Hubert & Arabie, 1985) values, which is a measure of the similarities between clusters, here the manually curated gold clusters and the clusters formed by DUREl. We clustered all graphs with different thresholds in the range of (0.3 0.325 0.35 ... 0.7) and selected the threshold with highest ARI for further analysis and experiments, resulting in a threshold of 0.6 for clustering.

Table 1: Cluster evaluation based on ARI

Headword	ARI	Headword	ARI
enkelspårig ('one-track, simplistic')	1.0	kriga ('make war')	0.614
fasad ('facade')	1.0	rutten ('rotten')	0.299
ofantlig ('immense')	1.0	ventilera ('ventilate')	0.291
baksida ('back')	0.863	lirka ('tinker, coax')	0.251
fotavtryck ('footprint')	0.84	vansinnig ('insane')	0.237
klimat ('climate')	0.772	hagla ('fall hail')	0.228
bagage ('baggage')	0.758	skör ('fragile')	0.068
vissen ('withered')	0.645	hemmaplan ('setting')	0.0
tvärnita ('jam on the breaks')	0.642	kapitulera ('capitulate')	-0.008
Average	0.528		

ARI generally ranges from -0.5 to 1. Values closer to 1 signify better agreement between the gold clusters and the derived clusters. A value of 0 suggests a random clustering while negative values indicate dissimilarity.

As can be noted, words such as the already mentioned *baksida* ('back, downside, drawback') and *enkelspårig* ('one-track, simplistic, narrow-minded') but also the noun *fasad* ('facade'), the adjective *ofantlig* ('immense'), and the noun *fotavtryck* ('footprint') have relatively high ARI values, indicating stronger alignment between their respective gold clusters and derived clusters. In contrast, the adjective *vansinnig* ('insane') and the verbs *hagla* ('hail') and *kapitulera* ('capitulate') show ARI values of 0 or slightly negative. An average of 0.528 across all words represent a reasonable overall similarity, and the results suggests that the automatically derived clusters generally encode meaningful semantic information which is useful for lexicographers.

In addition to the automatic ARI evaluation, the DUREl clusters were analyzed qualitatively by the lexicographers, which provided useful insights. For example, an examination of the figurative examples of *bagage* ('luggage') revealed that some of the uses constitute different variants of the idiomatic expression, *ha något i bagaget* ('to have something in the luggage') (see Schlechtweg et al., 2024). The idiom is not covered in the current edition of SO and will be included in the next edition.

Another example is the verb *hagla* (lit. 'fall hail'). In SO, the headword has two senses, the main sense 'fall hail' and the figurative subsense 'appear (over someone) in large quantity // mostly about abstract phenomena'. As shown in Figure 3, DUREl predicted three main semantic clusters which might be an indication that the word has three senses.

Among the blue-marked uses in Figure 3, the main weather sense of the verb in SO is represented. However, among the uses marked with an orange dot, there are sentences like "De [ungdomarna] står ofta på avstånd och kastar och för polisen blir det då att springa mot dem medan stenar haglar." ('They [the young people] often

stand at a distance and throw, and the police then have to run towards them while stones are hailing'.) Furthermore, among the green-marked dots there are samples like "Smädeorden haglade" ('the vituperations rained/came thick and fast'). In this case the SO lexicographers also concluded that, based on the editorial principles, the current subsense in the entry *hagla* is too wide. It should be split into two subsenses, one based on meaning extension concerning more concrete objects (like stones etc.), and the other one figurative, concerning abstract phenomena like vituperations. In other words, even if the correspondence between the gold clusters and the derived clusters is not perfect (as in the case of *hagla*; see Table 1 above), the lexicographers can get useful information about the meanings of a headword by using DUREl.

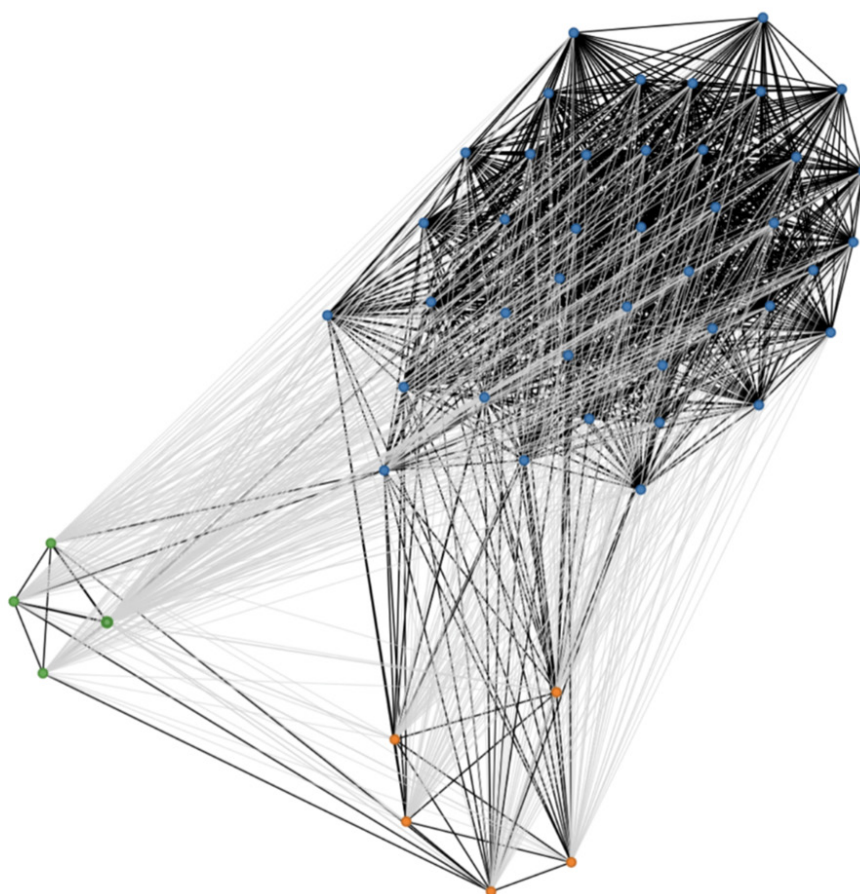


Fig. 3: DUREl sense cluster for the verb *hagla* ('fall hail; appear (over someone) in large quantity // mostly about abstract phenomena')

4.2 Experiment 2

After getting promising results from experiment 1, we started the second experiment with the major aim mentioned above. The methodology was the same, but this time we focused on the number of senses recorded in SO and compared that number with the number of clusters inferred for a given headword in the corpus using DUREl. A discrepancy in the number of senses and the recorded clusters can be an indicator that the dictionary article is outdated. Such headwords can be prioritized in order to update the dictionary.

During the first phase, we concentrated on the 39,000 nouns, adjectives and verbs with only one sense recorded in SO. A cursory review of this extensive data by the lexicographers showed that many headwords probably have more senses than the ones listed in the dictionary. An example is the noun *botemedel* which according to SO means “medel för att undanröja ett problem // ofta sjukdom men även allmänare” (‘means to eliminate a problem // often illness but also more generally’). The headword *botemedel* should, in line with current principles for meaning descriptions in the dictionary, be split into one main sense and one subsense (since the word is also used in more general contexts) (see Ralph et al., 1977; Järborg, 1989). The medical-related main sense could be illustrated by corpus-based examples including expressions like “jakten på ett botemedel mot AIDS” (‘the search for a cure for AIDS’) and the more general subsense by “Finns det ett botemedel mot trångsynthet och idioti?” (‘Is there a cure for bigotry and idiocy?’). A search in the SO database shows that there are about 180 other headwords in the dictionary with senses that should be split in the same way. In other words, based on individual words in the list with monosemic headwords, the lexicographers have found groups of headwords in SO whose semantic description could be improved.

Since DUREl has not been previously tested at that large scale, for a quick proof of concept we selected a set of random 281 headwords (nouns, adjective and verbs) in SO with only one sense registered. All headwords have at least 25 occurrences in the SVT corpus, which is an important prerequisite for being able to notice semantic patterns of the words.

Next, usages were extracted from the corpus and clusters were inferred as explained previously (see section 3) using the threshold parameter tuned in experiment 1. It appeared that 215 out of 281 words were predicted to have only one cluster (and hence one sense). Among those headwords you find nouns such as *brevinkast* (‘letterbox’) and *tisdag* (‘Tuesday’), adjectives like *oåterkallelig* (‘irreversible’) and *tunisisk* (‘Tunisian’) and verbs like *orsaka* (‘cause’) and *sukta* (‘long (in vain)’). According to the SO lexicographers’ assessment, these headwords, like many others in the list, have only one sense. The semantic description of those headwords in SO is thus satisfactory.

However, 49 of the 281 headwords in the experiment have two clusters. Furthermore, 9 have three clusters, 6 have four clusters, 1 has five clusters, and 1 has six clusters in DUREl. In other words, according to the annotation tool, these headwords were predicted to have between 2 to 6 senses.

While a full-scale proper evaluation of the results in this experiment still needs to be carried out, we were able to find promising cases by looking at some of the examples with more than one cluster. One example is the noun *lydnad* (‘obedience’) for which DUREl detected two clusters, colored blue and orange in Figure 4.

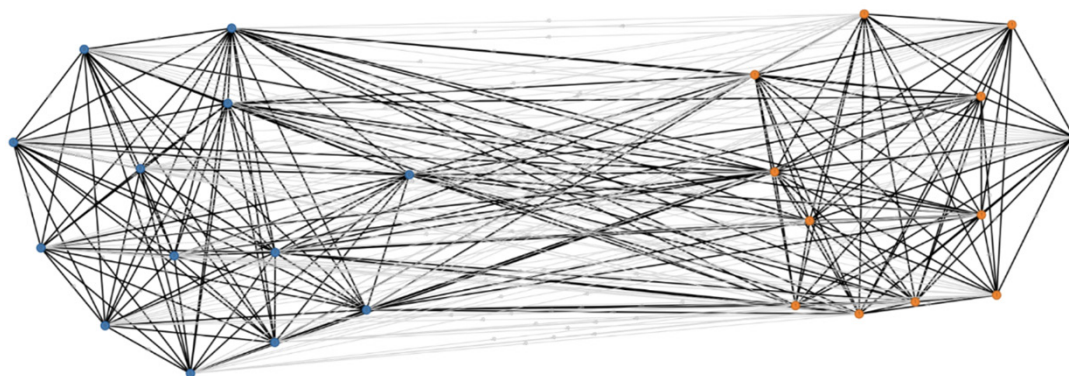


Fig. 4: DUREL sense cluster for the noun *lydnad* ('obedience').

The sense cluster by DUREL indicates that the noun *lydnad* has at least two different senses which was confirmed by the lexicographers who inspected the clusters and the corpus samples manually. They found that the cluster made up of orange dots includes corpus samples such as "Nunnor följer de tre klosterlöftena: att leva i celibat, fattigdom och lydnad. ('Nuns follow the three monastic vows: to live in celibacy, poverty and obedience.')

This example illustrates the meaning already recorded in SO. The cluster with blue dots, on the other hand, includes samples with word combinations like "tävla i lydnad" ('compete in obedience') and "nordiska mästerskap i lydnad" ('Nordic championships in obedience'). They represent a specialization of the sense which is related to dog- and equestrian sports. This specific meaning of the noun is not recorded in the current edition of SO and will be added in the next update.

Among the headwords with only one sense in SO but with two or more confirmed semantic clusters in the analysis by DUREL one may also find the noun *slutspurt* ('final spurt' but also 'finish'), *brottningsmatch* ('wrestling-match') and *avstickare* ('detour' but also 'digression'). Through the tool, the SO lexicographers have, among other things, been drawn to the figurative meanings of the three headwords in the corpus, which will result in improved dictionary articles.

Another example is the noun *nedtagning* with no less than five automatically generated semantic clusters by DUREL. However, according to SO, the word has just one sense: 'dampening the movement of a (high) ball with the body (chest) or leg // and placing it in front of the feet; in football'. In this case, it is not a subsense that is missing – the existing one is already very specific – and the lexicographers agree that a main, more basic, sense, which has to do with the act of 'taking down', should be added. The main sense of the word could be illustrated by corpus-based examples like "nedtagning av träd" ('removal of trees') och "uppsättning och nedtagning av belysning" ('setting up and taking down lighting').

However, based on the results of this experiment, it is also clear that the tool can be improved. For example, when it comes to nouns like *fransos* ('Frenchman') and *kulstötare* ('shot-putter'), DUREL shows two clusters, but the words have only one sense in SO and in the corpora. Furthermore, the headword *minfält* has just one cluster in DUREL, but a

closer look at the samples by the SO lexicographers shows that the word has two senses in Swedish: ‘minefield’ and ‘delicate subject’. The figurative meaning of the word is still missing in SO, so the dictionary article should be supplemented.

5. Results and Final Discussion

In this paper we presented two experiments. The aim of the first was to validate the feasibility and prove the usefulness of the annotation tool DUREl and its underlying computational language models for semantic variation and change detection. The aim of the second one was to use DUREl to revise semantic descriptions of headwords in SO. To sum up, a discrepancy between the number of senses recorded in SO and the number of clusters identified by DUREl can be used as an indicator that the description in the dictionary is outdated. Such headwords can be prioritized for manual inspection in order to update the dictionary entry.

The experiments are still small-scaled, but the results are nevertheless promising. So far, the work has, among other things, brought the lexicographers’ attention to subsenses (mostly figurative uses of the headwords but also meaning extensions and specializations) that should be added to the dictionary. Also missing main senses have been discovered. The work has also demonstrated that certain meaning descriptions in SO are too general and should be split in accordance with the current principles for the semantic descriptions in the dictionary. So far, we have mainly found well-established senses that should have been included in the latest edition of SO, but we have also found completely new senses in Swedish.

In this study we only used relatively modern Swedish corpora, and hence, we first of all registered lexical semantic variation. One way of capturing lexical semantic changes is by comparing semantic clusters based on corpora from different periods of time. If DUREl detects different numbers of clusters for a given word in different corpora, it may indicate that the word’s semantics have changed over time.

The study also provided new insights to improve the computational model and DUREl. Throughout our study we considerably improved the potential of DUREl to perform large-scale computational predictions and to visualize and manually inspect these. The lexicographers’ analysis of the clustered graphs can inform the future development of computational models, e.g., by retraining models on the errors identified by lexicographers.

The computer-aided methods are expected to streamline and enhance the lexicographic work on semantic variation and change detection. In the long run, this will help lexicographers to revise SO and other dictionaries in an efficient way.

References

Atkins, B. T. S., & Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford University Press.

Blensenius, K., Sköldberg, E., & Bäckerud, E. (2021). Finding gaps in semantic descriptions. Visualisation of the cross-reference network in a Swedish monolingual dictionary. In I. Kosem, M. Cukr, M. Jakubiček, J. Kallas, S. Krek, S., & C. Tiberius (Eds.), *Electronic lexicography in the 21st century. Proceedings of the eLex 2021 conference. 5–7 July 2021* (pp. 247–258). Online. Brno.

Borin, L., Forsberg, M., & Roxendal, J. (2012). Korp: The corpus infrastructure of Språkbanken. In N. Calzolari, K. Choukri, T. Declerck, M. Uğur Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of LREC 2012* (pp. 474–478). Istanbul: ELRA.

Cassotti, P., Siciliani, L., de Gemmis, M., Semeraro, G., & Basile, P. (2023). XI-lexeme: Wic pretrained model for cross-lingual lexical semantic change. In *Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics*. (Volume 2: Short Papers) (pp. 1577–1583). Online. Association for Computational Linguistics.

Cavallin, K. (2012). Exploring semantic change with lexical sets. In R. V. Fjeld, & J. M. Torjusen (Eds.), *Proceedings of the XV EURALEX international congress* (pp. 1018–1022). Oslo.

Cook, P., Lau, J. H., Rundell, M., McCarthy, D., & Baldwin, T. (2013). A lexicographic appraisal of an automatic approach for detecting new wordsenses. In I. Kosem, J. Kallas, P. Gantar, S. Krek, M. Langemets, & M. Tuulik (Eds.), *Proceedings of eLex 2013 conference* (pp. 49–65). Ljubljana/Tallinn: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2, 193–218.

Järborg, J. (1989). *Betydelseanalys och betydelsebeskrivning i Lexikalisk databas. Preliminär version*. Göteborg: Göteborgs universitet.

Korp: Språkbanken's word research platform. Retrieved May 15, 2024, from <https://spraakbanken.gu.se/korp>

Lau, J. H., Cook, P., McCarthy, D., Newman, D., & Baldwin, T. (2012). Word sense induction for novel sense detection. In W. Daelemans (Ed.), *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 591–601). Avignon, France: Association for Computational Linguistics.

Lew, R. (2013). Identifying, ordering and defining senses. In H. Jackson (Ed.), *The Bloomsbury companion to lexicography* (pp. 284–302). Bloomsbury.

Malmgren, S.-G., & Sköldberg, E. (2013). The Lexicography of Swedish and other Scandinavian Languages. *International Journal of Lexicography*, 26(2), 117–134. <http://dx.doi.org/10.1093/ijl/ect008>

Nimb, S., Hartvig Sørensen, N., & Lorentzen, H. (2020). Updating the dictionary: Semantic change identification based on change in bigrams over time. *Slovenščina* 2.0 8(2), 112–138. <https://doi.org/10.4312/slo2.0.2020.2.112-138>

Petersson, S., & Sköldberg, E. (2021). Semantic change in Swedish – from a lexicographic perspective. In N. Tahmasebi, L. Borin, A. Jatowt, Y. Xu, & S. Hengchen (Eds.), *Computational approaches to semantic change* (pp. 149–167). Language Science Press.

Ralph, B., Järborg, J., & Allén, S. (1977). *Svensk ordbok och Lexikalisk databas. Förstudierapport*. Göteborg: Göteborgs universitet.

Schlechtweg, D., Schulte im Walde, S., & Eckmann, S. (2018). Diachronic Usage Relatedness (DUREl): A framework for the annotation of lexical semantic change. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 169–174). New Orleans, Louisiana.

Schlechtweg, D., McGillivray, B., Hengchen, S., Dubossarsky, H., & Tahmasebi, N. (2020). SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation* (pp. 1–23). Barcelona (online), International Committee for Computational Linguistics.

Schlechtweg, D., Virk, S. M., Sander, P., Sköldberg, E., Theuer Linke, L., Zhang, T., Tahmasebi, N., Kuhn, J., & Schulte im Walde, P. (2024). The DUREl Annotation Tool: Human and Computational Measurement of Semantic Proximity, Sense Clusters and Semantic Change. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations* (pp. 137–149). St. Julians, Malta: Association for Computational Linguistics.

SO 2021: *Svensk ordbok utgiven av Svenska Akademien* [‘The Contemporary Dictionary of the Swedish Academy’]. (2021). 2nd edition. Retrieved May 15, 2024, from <https://svenska.se/>

Svensén, B. (2009). *A handbook of lexicography: The theory and practice of dictionary-making*. Cambridge University Press.

SVT (Sweden Television) corpus. Retrieved May 15, 2024, from Språkbanken’s word research platform Korp <https://spraakbanken.gu.se/korp>

Taher Pilehvar, M. & Camacho-Collados, J. (2019). WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Volume 1 (Long and Short Papers) (pp. 1267–1273). Minneapolis, Minnesota: Association for Computational Linguistics.

Acknowledgements

This work has jointly been funded by the research program Change is Key! supported by Riksbankens Jubileumsfond (under reference number M21-0021) and the Swedish Academy.

Contact information

Emma Sköldberg

University of Gothenburg, Department of Swedish, multilingualism, language technology
emma.skoldberg@svenska.gu.se

Shafqat Mumtaz Virk

University of Gothenburg, Department of Swedish, multilingualism, language technology
shafqat.virk@svenska.gu.se

Pauline Sander

University of Stuttgart, Institute for Natural Language Processing
pauline.sander@ims.uni-stuttgart.de

Simon Hengchen

iguanodon.ai and Université de Genève
simon@iguanodon.ai

Dominik Schlechtweg

University of Stuttgart, Institute for Natural Language Processing
dominik.schlechtweg@ims.uni-stuttgart.de

Petra Storjohann

SYNONYMS IN CONTRAST

A Dynamic and Descriptive Resource for New Semantic Equivalents

Abstract This paper introduces “Synonyms in Contrast”, a new online dictionary that addresses the complexities and nuances of neologistic (near-) synonyms in the German language. The emergence of new lexical items, often borrowings from English, has contributed to the proliferation of meaning equivalents. These share a large extent of contextual features, causing ambiguity and uncertainty among speakers regarding their appropriate usage. The dictionary, which is part of the new IDS *Neo*²⁰²⁰⁺ resource, distinguishes itself by utilising corpus-driven and corpus-based approached and sophisticated lexicographic methods such as word embeddings to provide users with comprehensive, context-sensitive information on semantic similarities and differences between new synonyms. This resource aims to aid in the clear understanding of new terms, their usage, and interrelations by detailing their semantic overlaps, distinctive contexts, and collocational preferences. Through a combination of empirical data and user-specified presentation, the dictionary facilitates better linguistic decisions in both everyday and also more specialised communication. The ultimate goal is to enhance the integration of new lexical variants into the mental lexicon of German speakers by offering a reliable, dynamic, and descriptive platform for exploring and resolving lexical uncertainties with respect to meaning equivalents that have recently emerged.

Keywords dictionary of synonyms; discriminating meaning equivalents; contrastive lexicography; semantic similarities; new lexical variants

1. Introduction

Parts of the new German vocabulary are lexical items that share similar meanings (synonyms in their broadest sense)¹, potentially leading to ambiguity in certain linguistic contexts or to uncertainty as to their preferences and appropriateness. This can be attributed to their unfamiliarity, semantic overlaps, conceptual proximity or it can also be a result of incomplete lexicalization processes. Some newly introduced terms are adopted from English (*Lockdown/Shutdown*). Other English terms are borrowed and serve as equivalent counterparts to existing native words (*Fake-News/Falschnachrichten/Fake-Nachrichten/Fakemeldung/Falschmeldung*, *Self-Care/Selbstfürsorge*, *Prank/Streich*). Alternatively, they enter concurrently together with their loan translations (*Wallbox/(Wand)Ladestation*, *Greenflation/grüne Inflation/Grünflation*). In situations where clusters of synonymous neologisms or a number of similar morphological variants emerge at the same time, users are presented with multiple options for expressing similar concepts and may feel the need to distinguish between them to determine their appropriate use. These challenges among German

¹ For a detailed discussion on classifications of synonymy see Cruse (1986), for a more pragmatic approach see Murphy (2003; 2013).

speakers become apparent in online forums, as attested by questions such as “What does “Prank” (*prank*) mean and what is the difference to a “Streich” (*hoax*)?”

This paper introduces a new dictionary for new German synonyms “Synonyme im Kontrast” (‘Synonyms in Contrast’), including corpus-driven and corpus-based methods (cf. Tognini-Bonelli, 2001) and addresses both, meta-lexicographic reflections and the practical implementation. So far, there is no corpus-related tradition in identifying and analysing synonyms, including neo-synonyms. Consequently, this gap has repercussions on the lexicographic documentation of terms with similar meanings. This absence has led to a lack of innovative reference guides that employ new approaches to detecting, describing and presenting synonyms. This is noteworthy considering because German lexicography has a well-established tradition of synonym dictionaries (cf. Hausmann, 1990; Hahn, 2004), and present-day opportunities allow for an empirically grounded online lexicography. As Murphy (2013, p. 301) points out “we can be very imaginative in how synonym information is found and presented”. As corpus-guided² reference works are inclined to follow descriptive perspectives, their aims differ. In contrast to distinctive or cumulative reference works (Hahn, 2004), a descriptive synonym resource aims to provide terms that are semantically similar or equivalent, along with comprehensive comparative explanations. The headwords are thoroughly disambiguated, highlighting context-bound similarities and differences, supported by authentic examples.

2. Neologistic Synonyms in Contrastive Entries

The new synonym guide “Synonyme im Kontrast” is a fundamental integral component of the recently designed neologism resource IDS *Neo*²⁰²⁰⁺ which is currently being implemented.³ A central feature of the synonym guide is its elaborate contrastive entries which allow for context-bound comparison of lexical use. Following the main principles developed for the German paronym dictionary for easily confused words (“Paronyme – Dynamisch im Kontrast”), the new synonym resource continues to establish a lexicography of lexical-semantic doubt. In order to retrieve and accommodate for necessary data, complementary corpus-related approaches are employed. On the one hand, potential synonyms are being extracted automatically through the comparison of similar collocation profiles, thus using a corpus-driven method. On the other hand, both near-synonym terms are subsequently searched for in close contextual vicinity in the underlying corpus. Then, the lexicographer can verify their semantic relation using the corpus-based method.

2.1 Lexicographic Corpus Methods

New methods for retrieving synonyms include word embedding tools, which use vector-based analysis to identify paradigmatic commonalities. These tools can identify

² In this paper, the term ‘corpus-guided’ refers to any method related to corpus-linguistic work.

³ IDS *Neo*²⁰²⁰⁺ will comprise a neologism dictionary (“Neo Dictionary”) launching in 2024; a dashboard featuring various formats and details on neologism research (“Neo Dashboard”) and a neologistic synonym guide (“Synonyms in Contrast”) launching in 2025.

#	S _{cos}	similar by w2v
1	1.000	Lockdown
2	0.920	Corona-Lockdown
3	0.900	Shutdown
4	0.888	Teil-Lockdown
5	0.858	Corona-Shutdown
6	0.825	Lockdowns
7	0.761	Corona-Winter
8	0.755	Corona-Lockdowns
9	0.753	Shutdowns
10	0.744	Pandemie
11	0.735	Corona-Einschränkung
12	0.733	Lockdown-Maßnahme
13	0.727	Lockdown-Phase
14	0.723	Teil-Lockdowns
15	0.722	Kontaktbeschränkung
16	0.721	Corona-Restriktion
17	0.714	Corona-Beschränkung
18	0.712	Kontaktsperrn
19	0.711	Kontaktbeschränkung
20	0.708	Coronakrise
21	0.704	Ladenschließungen

Each word listed provides an interactive link to generate the corresponding word embedding accordingly. To answer user questions about the exact differences between two terms, one must employ a complementary method and conduct a collocation analysis. Collocations and examples that demonstrate contextual usage serve as more than just means of lexicographic analysis and interpretation; even more so they constitute a vital component within each entry, requiring users to consult them for a comprehensive understanding. Since the early stages of corpus-guided lexicography, as advocated by Partington (1998), the importance of this approach has been recognized and continues to be emphasised.

The clear demonstration of the precise circumstances in which a new lexical item can be substituted for a similar item, by including typical syntagmatic patterns,

empowers users. This approach helps them to appropriately connect new words with similar meanings to the unknown concepts they represent and to understand the relationships between the lexical items (see Table 1). The question then arises: How can such pairs or groups be semantically registered and documented in a dictionary? Providing a parallel description of patterns in an entry enables users to discern where differences lie, or whether both terms are interchangeable.

Table 1: Typical collocates allocated to context, word class and semantic group

	<i>Lockdown</i>	<i>Shutdown</i>
adjectives		
How is X being characterised?	<i>hart</i> ('hard'), <i>light</i> ('light'), <i>erneut</i> ('anew/again'), <i>total</i> ('total'), <i>partiell</i> ('partial'), <i>monatelang</i> ('for months'), <i>flächendeckend</i> ('nationwide'), <i>coronabedingt</i> ('covid-related'), <i>strikt</i> ('strict'), <i>neuerlich</i> ('repeated')	<i>hart</i> ('hard'), <i>light</i> ('light'), <i>erneut</i> ('anew'), <i>komplett</i> ('complete'), <i>teilweise</i> ('partial'), <i>erschöpfend</i> ('tiring'), <i>coronabedingt</i> ('Covid-related'), <i>wochenlang</i> ('for weeks'), <i>verschärft</i> ('tightened'), <i>zielführend</i> ('effective')
verbs		
What does someone do in X?	<i>überstehen</i> ('endure'), <i>leiden</i> ('suffer'), <i>nutzen</i> ('use'), <i>vermissen</i> ('miss'), <i>entschleunigen</i> ('slow down'), <i>weitermachen</i> ('continue'), <i>verlängern</i> ('extend'), <i>verschärfen</i> ('intensify'), <i>verhindern</i> ('prevent'), <i>lockern</i> ('loosen')	<i>überstehen</i> ('endure'), <i>überleben</i> ('survive'), <i>aushalten</i> ('endure'), <i>abgewinnen</i> ('gain'), <i>beschließen</i> ('agree'), <i>verhindern</i> ('prevent'), <i>fordern</i> ('demand'), <i>verlängern</i> ('extend'), <i>lockern</i> ('loosen'), <i>verschärfen</i> ('intensify')
What does X do ?	<i>bremsen</i> ('slow down'), <i>(an-)dauern</i> ('last'), <i>gelten</i> ('apply'), <i>enden</i> ('end'), <i>herrschen</i> ('dominate'), <i>kommen</i> ('come'), <i>beginnen</i> ('begin'), <i>verhindern</i> ('prevent'), <i>drohen</i> ('threaten')	<i>(an-)dauern</i> ('last'), <i>zwingen</i> ('force'), <i>bremsen</i> ('slow down'), <i>drohen</i> ('threaten'), <i>treffen</i> ('affect'), <i>kommen</i> ('come'), <i>verhindern</i> ('prevent'), <i>enden</i> ('end'), <i>kosten</i> ('cost'), <i>eintreten</i> ('happen')
nouns		
Who is affected by X?	<i>Ungeimpfte</i> ('unvaccinated'), <i>Gastronomie</i> ('gastronomy'), <i>Geimpfte</i> ('vaccinated'), <i>Schulen</i> ('schools'), <i>Einzelhandel</i> ('retail'), <i>Wirtschaft</i> ('economy'), <i>Geschäfte</i> ('shops'), <i>Restaurant</i> ('restaurants'), <i>Friseure</i> ('hairdressers')	<i>Wirtschaft</i> ('economy'), <i>Möbelindustrie</i> ('furniture industry'), <i>Einzelhandel</i> ('retail'), <i>Selbstständige</i> ('self-employed'), <i>Restaurants</i> ('restaurants'), <i>Onlinehändler</i> ('online retailer'), <i>Kulturinstitutionen</i> ('cultural institutions'), <i>Bürger</i> ('citizens'), <i>Ungeimpfte</i> ('unvaccinated'), <i>Gastronomie</i> ('gastronomy')
What is X being associated with X?	<i>Coronapandemie</i> ('Covid pandemic'), <i>Homeoffice</i> ('homeoffice'), <i>Isolation</i> ('isolation'), <i>Ausgangs- und Kontaktbeschränkungen</i> ('curfews and contact restrictions'), <i>(Corona)-Infektionszahlen</i> ('Covid infection numbers'), <i>Öffnungsperspektive</i> ('opening perspective'), <i>Impfstrategie</i> ('vaccination strategy'), <i>Hamsterkäufe</i> ('panic buying'), <i>Impfpflicht</i> ('compulsory vaccination'), <i>Home schooling</i> ('home schooling')	<i>Reisesperren</i> ('travel bans'), <i>Lockerungen</i> ('easing'), <i>Corona-Pandemie</i> ('Covid pandemic'), <i>Homeoffice</i> ('home office'), <i>Infektionszahlen</i> ('infection numbers'), <i>Einschränkungen</i> ('restrictions'), <i>Grundrechte</i> ('basic rights'), <i>Home schooling</i> ('home schooling'), <i>Social Distancing</i> ('social distancing'), <i>Eindämmung</i> ('containment')

Table 1 shows that both lexical items overlap semantically to a large degree. Context-bound semantic commonalities as well as differences discernible through the interpretation of collocates, thereby establish the analytical foundation for frame-based referential use, as registered in the dictionary. A semantically equivalent relationship is also confirmed by a number of corpus examples (see example (1)).

Die Wissenschaftler hatten mehrere Szenarien entwickelt. Sie erwarten demnach, dass bei einem zweimonatigen **Shutdown** die Emissionen bundesweit um etwa 40 Millionen Tonnen CO₂ gegenüber 2019 sinken. Bei einem zweiten **Lockdown** aufgrund einer zweiten Infektionswelle wären Einsparungen von weiteren rund 71 Millionen Tonnen CO₂ möglich. (DeReKo, Welt Online, 25.05.2020) (Scientists had developed several scenarios. Accordingly, they expect that a two-month shutdown would reduce emissions nationwide by about 40 million tons of CO₂ compared to 2019. With a second lockdown due to a second wave of infections, additional savings of about 71 million tons of CO₂ could be possible.)

2.2 Two-level Presentation of Similarities and Differences

The typical co-occurrences in Table 1 and the corpus examples suggest a synonymous relationship between *Lockdown* and *Shutdown*. However, this relationship is strictly context-bound, as *Shutdown* is a polysemous term with two older senses not attested for *Lockdown*. Consequently, these neologisms also exhibit differences in their usages. It is therefore important to clearly indicate both semantic overlaps and differences, as identifiable through the use of collocates in a dictionary entry (Figure 2). In an entry overview, all senses of the terms in question are listed, with identical contextual uses appearing first by default. However, other listing options are available via a top menu. Each sense (or tile) contains a sense identifier in the form of a synonym or hyperonym, a domain and exemplary collocates.

The screenshot shows the IDS Neo 2020+ interface. At the top, it says 'Synonyme im Kontrast'. Below this, there are filters for 'freq' and 'sortiert nach Gemeinsamkeiten'. The main content area displays two terms: 'Lockdown (Neulexem)' and 'Shutdown (Neubedeutung)'. Each term has a 'Stillegung' (Stylization) section with a list of collocates (e.g., 'Ungeimpfte, Gastronomie, Geimpfte, Schulen, Einzelhandel ...') and a 'mehr' link. The 'Shutdown' section also includes 'Haushaltssperre' and 'Herunterfahren' with their respective collocates and 'mehr' links.

Fig. 2: Contrastive overview of *Lockdown/Shutdown* with all their senses

Senses that are not shared are also marked by different colours and positions. Each use can be further explored by clicking on it individually or by selecting up to three tiles for comparison (Figure 3). In Figure 3, differing senses such as *Lockdown* ‘Stilllegung’ (‘closing down’) and *Shutdown* ‘Haushaltssperre’ (‘budget freeze’) were chosen for the detailed view to illustrate context-bound collocational differences. Here, an explicit definition is followed by details on domain and subject, a list of grouped collocates and background information. All details are presented side by side for direct comparison.

Verwendung	Grammatik	Verwendung	Grammatik
Lockdown, ‘Stilllegung’ (Neulexem: seit 2020) bezeichnet einen Zeitraum, in dem fast alle wirtschaftlichen und gesellschaftlichen Aktivitäten auf politische Anordnung hin stillgelegt sind (z.B. zum Infektionsschutz) Domäne: meist in GESUNDHEIT, GESELLSCHAFT, POLITIK Thematischer Bezug: BESONDERS WÄHREND CORONAPANDEMIE Kollokationen Wer oder was ist von einem Lockdown betroffen? <i>Ungeimpfte, Gastronomie, Geimpfte, Schulen, ...</i> Was macht man in bzw. mit einem Lockdown? <i>überstehen, leiden, nutzen, vermissen, entschleunigen, ...</i> Was macht ein Lockdown? <i>bremsen, dauern, gelten, enden, losgehen, ...</i> Wie ist ein Lockdown? <i>hart, light, erneut, total, partiell, monatelang, ...</i> Was wird im Zusammenhang mit Lockdown thematisiert? <i>Coronapandemie, Homeoffice, Isolation, Ausgangssperren, ...</i>		Shutdown, ‘Haushaltssperre’ bezeichnet eine Haushaltssperre, die zu einem Teil-Stillstand der Regierungsgeschäfte in den USA führt und daraufhin die Tätigkeiten des Staats- und Verwaltungsapparates einstellt. Domäne: meist in POLITIK Thematischer Bezug: US-Regierung Kollokationen Wer oder was ist von einem Shutdown betroffen? <i>US-Kongress, US-Wirtschaft, Regierungsgeschäfte, US-Regierung, ...</i> Was macht man in bzw. mit einem Shutdown? <i>abwenden, verhindern, beenden, vermeiden, erzwingen, ...</i> Was macht ein Shutdown? <i>andauern, anhalten, treffen, lähmen, kommen, verhindern, ...</i> Wie ist ein Shutdown? <i>erneut, teilweise, amerikanisch, neuerlich, wochenlang, ...</i> Was wird im Zusammenhang mit Shutdown thematisiert? <i>Fristablauf, Stilllegung, Trump, US-Geschichte, Demokraten, Republikaner, ...</i>	
Hintergrund: Der Ausdruck <i>Lockdown</i> bezieht sich im Zuge der Corona-Pandemie auf einen Ausnahmezustand und Zeitraum, in dem zahlreiche Bereiche des öffentlichen Lebens stillgelegt waren. Der Ausdruck gehört damit zum Coronawortschatz und war besonders zwischen 2020-2022 besonders präsent.		Hintergrund: Bei <i>Shutdown</i> handelt es sich möglicherweise um eine Kurzform von Government-Shutdown und der Ausdruck bezieht sich auf US-amerikanische politische Zustände, die erfolgen, wenn sich Senat, Repräsentantenhaus und Präsident nicht rechtzeitig über die Bewilligung von Haushaltsmitteln einigen. <i>Shutdowns</i> ereignen sich regelmäßig.	

Fig. 3: Contrastive detail view of *Lockdown/Shutdown* and two differing senses

The dictionary includes details specifically pertinent to new words, such as the type of neologism (e.g., loan word/translation, newly coined through word formation, see Figure 2), time of introduction and extent of its integration (e.g., as seen in grammar paradigms, see Figure 3).

3. Summary

The new German dictionary “Synonyme im Kontrast” will include neologistic near-synonyms of various kinds. Firstly, it will comprise pairs in contrastive entries consisting of one neologistic term and one older, often native term. Secondly, one will find entries with two new borrowed words that have entered the German language simultaneously. Thirdly, there will also be cases where a new loanword emerged and was given a German loan translation at the same time. In some cases, synonyms act as lexical rivals, displaying clear preferences in usage. In other instances, they coexist, leading to situations of lexical uncertainty where users are unsure about which term to use. The new reference guide aims at answering questions commonly found in online forums regarding the use of two new, semantically similar words.

Successful storage of information in the mental lexicon can only be achieved when linguistic and encyclopaedic knowledge are intricately intertwined within an entry. This can only be guaranteed when information related to two headwords is studied together, addressing both semantic and syntactic questions, including collocational and colligational issues. The explanations are supported by examples extracted from corpus data, with entries that aim to advance lexicographic knowledge in the realm of synonymy.

References

- Cruse, A. (1986). *Lexical Semantics*. Cambridge University Press.
- DeReKo: *Deutsches Referenzkorpus*. Retrieved May 1, 2024, from <https://cosmas2.ids-mannheim.de/cosmas2-web/>.
- DEREKOVECS. Retrieved May 1, 2024, from <https://corpora.ids-mannheim.de/derekovecs>
- gutefrage.net*. Retrieved May 1, 2024, from <https://www.gutefrage.net/>
- Hahn, M. (2004). Synonyme und Quasisynonyme in lexikografischer Hinsicht. *Neuphilologische Mitteilungen*, 105(1), 3–20.
- Hausmann, F. J. (1990). The Dictionary of Synonyms: Discriminating Synonymy. In F. J. Hausmann, O. Reichmann, H. E. Wiegand & L. Zgusta, L. (Eds.), *Wörterbücher. Dictionaries. Dictionnaires*. HSK 5.2. (pp. 1067–1075). De Gruyter.
- Murphy, M. L. (2003). *Semantic Relations and the Lexicon*. Cambridge University Press.
- Murphy, M. L. (2013). What we talk about when we talk about synonyms. *International Journal of Lexicography*, 26(3), 279–304. <https://doi.org/10.1093/ijl/ect023>
- Paronyme – Dynamisch im Kontrast*. Retrieved May 1, 2024, from <https://www.owid.de/parowb/>
- Partington, A. (1998). *Patterns and Meanings*. Benjamins.
- Tognini-Bonelli, E. (2002). *Corpus Linguistics at Work*. Benjamins.

Contact information

Petra Storjohann

Leibniz-Institut für Deutsche Sprache, Mannheim
storjohann@ids-mannheim.de

BACK TO BASICS: MEANING DESCRIPTION FOR HUMAN USERS AND FOR COMPUTERS

Abstract In this article, we return to a classic lexicographical topic and address some aspects involved in the practice of defining. Digital developments have increasingly required dictionary definitions to operate independently of others if they are to be utilised in new contexts, possibly even detached from the original dictionary presentation. We examine two types of definitions where the problem is particularly obvious: morpho-semantically related words and inherited information. The first type refers to words with the same semantic core that appear in regular morphological derivations, in this context illustrated by triplets such as *pluralism*, *pluralist*, *pluralistic*. The second type is to do with senses that presuppose the definition of an earlier, usually superordinate sense, as when a culinary sense is singled out: ‘this fish used as food’. Five dictionaries are compared and analysed to uncover the tendencies and strategies followed. All five dictionaries show internal inconsistency, and while there is general consensus that the traditional paper model should be abandoned, the dictionaries do not necessarily agree on strategy and solutions.

Keywords semantics; meaning description; data design; Scandinavian dictionaries; regular polysemy; inheritance; morpho-semantics

1. Introduction

Over the past decades, digital developments have led many dictionaries to transition from print to electronic formats. In the process, the underlying data have undergone several changes to adapt to the new format. To name a few, the system of reference has been replaced by hyperlinks, abbreviations are shown in full, audio files replace or supplement phonetic notation for pronunciation, and multiword expressions are presented under each of the constituent words. In this article, we examine some aspects of definitional practice that can be seen as leftovers from the paper dictionary but have not attracted as much attention as the ones just mentioned.

The title of the article should not be taken to mean that we aim to rethink every aspect of meaning description in dictionary definitions. The goal is to examine a few limited practices that are no longer entirely appropriate, given that lexicographic resources are increasingly stored in databases that can be used for multiple purposes, not all of which are necessarily lexicographic in nature. Examples include the DANTE project (cf. Atkins et al., 2010) and the multilingual datasets used by K Dictionaries (cf. Lonke et al., 2022). One consequence of digital developments is that dictionary entries may not always be presented in full, and users may not always see the nearest alphabetical neighbour entries. For example, a user may access the dictionary data via an API to get explanations of words in running text as a service provided by an e-reader. Ideally, the user wants to see one and only one explanation: the correct definition

in the specific look-up situation. Another development is the increased demand for semantic descriptions from language technology, such as computer programs that need semantic information about the words in texts to create an automatic topic classification.

If we want the dictionary data to be prepared for uses such as these, dictionary explanations must be phrased so as to function independently of other definitions. In other words, there are reasons to consider whether the current practice should be revised in this respect. In the dictionary under primary investigation, The Danish Dictionary, we have noted two aspects of definitional practice that have repeatedly proved problematic, and they are the focus of this study.

2. Morpho-Semantically Similar Words

The first aspect we will address concerns different but closely related headwords. Derivations of words with the same semantic core often appear as alphabetical neighbours, and their descriptions are traditionally handled in a similar way: one word is treated as central, while the others explain how they are linguistically related to the core sense without repeating the entire definition (cf. Svensén, 2009, pp. 227–230).

Typical examples of this include political ideologies and religious or philosophical beliefs, which are often realised as linguistic triplets: the system of thought is treated as the core element and expressed as *-ism*, a member or follower is formed by means of *-ist*, and a corresponding adjective is formed by means of *-istic*.

An example from The Danish Dictionary and the traditional template can be illustrated by the entries *Buddhism*, *Buddhist* and *Buddhistic* (the headwords and definitions are translated from the Danish original for ease of understanding):

1. *Buddhism*: an East and Central Asian religion that originated c. 500 BCE based on Buddha's teachings that life is full of sorrow and suffering, from which one can only be freed by renouncing greed, hatred and deceit
2. *Buddhist*: a person who believes in Buddhism
3. *Buddhistic*: relating or belonging to Buddhism

In a printed dictionary, this solution has the dual advantage of saving space while simultaneously showing the relationship between morphologically and etymologically similar words. However, if the derived entry appears independently of the base word, it is a different matter, and the user is likely to feel poorly informed by this model (cf. Atkins & Rundell (2008, p. 436): "As all user research shows, it is never a good idea to require users to go to a second (or third) entry to find the information they are looking for in a first").

3. Inheritance

In a printed dictionary, it is not uncommon to find a definition that presupposes the definition of another, usually superordinate sense. If the subsense simply emphasises one aspect of the general meaning, it may feel excessively cumbersome to repeat a potentially complicated explanation of the core sense (cf. discussion in Atkins & Rundell, 2008, p. 446). We call this phenomenon *inheritance*: aspects of the general sense are ‘inherited’ by other senses of the headword. Consider the following examples taken from The Danish Dictionary (as above translated from the Danish original for ease of understanding):

Ablative

- (1) a grammatical case for a noun or a noun phrase that typically expresses direction away from something, or circumstances such as place, means, manner or cause – found in languages such as Sanskrit, Latin and Finnish

1(a) a word form in **this** case

Pyramid

- (1) an Egyptian funerary monument, often of very large dimensions, built of stone and featuring a square base and four triangular sides that converge to a point at the top

1(a) a building or structure resembling **such** a funerary monument

From these examples, it is easy to see that when read detached from their contexts as aspects of a more elaborate description, the subsense definitions in 1(a) are not very useful.

Subsense definitions that follow this model are especially common with words that are prone to *regular polysemy*, which refers to the situation where a group of words within a semantic domain share a potential for ambiguity (cf. Svensén, 2009, p. 209). A case in point is *school*, which may refer to (a) a concrete building (‘the school was built of yellow bricks’), (b) the institution that is housed in the building (‘go to school’), or (c) the people who work there (‘the school decided to admit new students’). Potentially, all other words for buildings may behave like *school* in sharing these three meaning aspects. Similarly, many animals and plants can be described both as biological species and as terms for culinary items: ‘I love apples’, ‘put the chicken in the oven’ (cf. Sørensen et al., 2023). Other examples taken from The Danish Dictionary include the following:

Plaice

- (1) a 40–90 cm long flatfish with a grey-brown upper side adorned with red dots and a white underside

1(a) **this** fish used as food

Garlic

- (1) a plant with white flowers tinged with reddish or greenish hues, and white, strongly scented, and tasty bulbs that form a row of wedge-shaped side bulbs during growth

1(a) bulbs from **this** plant, used as food, herbal medicine, etc.

In the subsenses, one particular aspect is singled out whereas the core meaning itself is inherited from the superordinate sense and only referred to by means of an anaphoric expression (*this*, *such* etc.). This solution may work when both senses are presented and read consecutively, but is less obvious if this is not the case. And if the subsense is the only one presented to the user, it becomes completely useless (cf. Marellø, 1988).

4. Examination of Templates Used in Existing Dictionaries

Not all solutions chosen for the printed dictionary are suitable for the electronic medium and need to be reconsidered. This is something that we as editors of The Danish Dictionary have realised over the years but the practical task of revising all existing relevant entries is far from systematically carried out. Nor have we settled on exactly which solution to use instead. To help us determine this, we have carried out a comparative study involving five existing dictionaries of contemporary language: two Danish dictionaries, one Norwegian, one Swedish and one English dictionary. The first four are all dictionaries for Scandinavian languages and are well suited for comparison as the three languages are closely related. The English dictionary was included in the investigation in order to broaden the perspective. These dictionaries were selected as they all originate from printed books that have been transformed into online dictionaries and thus are likely to have gone through similar processes and considerations. The five dictionaries included in the study are as follows:

- The Danish Dictionary (*Den Danske Ordbog*, DDO): published in print 2003–2005, online since 2009
- The Danish Internet Dictionary (*Den Danske Netordbog*, DDNO): published in print under another title (*DanskOrdbogen*, 1999), revised and published online since 2003
- The Norwegian Academy Dictionary (*Det Norske Akademis Ordbok*, NAOB): published in print under another title (*Norsk Riksmålsordbok* 1937–1957, supplement 1995), retrodigitised, revised and published online since 2017
- The Swedish Dictionary (*Svensk ordbok*, SO): published in print 2009, online since 2015
- *Oxford Dictionary of English*, ODE (3 ed.): published in print 2010, online since 2015

For each of the two types of definitions, a number of lexical items were selected and analysed in order to uncover tendencies and strategies used by the five dictionaries. For type 1) the lexical items were made up of 33 triplets of the above kind, and for type 2) they were chosen as subsenses from words that could potentially be described by means of an anaphoric expression.

4.1 Investigating Morpho-Semantic Similarity

As mentioned in Section 2, many words occur in triplets following the pattern *-isme*, *-ist*, *-istisk* (in Danish), denoting political ideologies (*zionisme*, *zionist*, *zionistisk*), artistic movements (*fauvisme*, *fauvist*, *fauvistisk*), religious (*shintoisme*, *shintoist*, *shintoistisk*) and philosophical beliefs (*nihilisme*, *nihilist*, *nihilistisk*), or medical or psychological conditions (*autisme*, *autist*, *autistisk*), etc. For the investigation of this type, we selected 33 such triplets from The Danish Dictionary (DDO) and then compared their definitions with the corresponding words in the other dictionaries, thus potentially comprising nearly 500 definitions. However, not all the triplets had equivalents in the other dictionaries, something we did not know or check in advance.

A pre-analysis of the triplets from DDO revealed both consistencies and inconsistencies in the way they are treated. For all the triplets, the *-isme* element was used as the core element and explained by a full definition. In the majority of cases (two-thirds of the triplets), the solution adopted for the *-ist* and *-istisk* elements was the same: both follow the same template, i.e., one of the four described below.

Even in case the derived members of the triplets did not use the same solution, there was a tendency for them to follow one of four possible templates:

1. the traditional paper solution where *-ist* is defined as ‘a person who follows/ believes in *-ism*’ and *-istic* as ‘denoting or concerning *-ism*’
2. the same as (1), but where the *-ism* element is a clickable link
3. a partially unfolded definition where the *-ism* element is mentioned, and usually clickable, and also given a short explanation
4. a fully unfolded definition where each headword of the triplet has a full definition that can be read independently of the others and with no mention of the core element

The four templates are illustrated in Table 1 with examples taken from DDO (in our translation; text in <...> are hyperlinks).

Table 1: Templates used for the description of morpho-semantically similar words

	-ism, noun	-ist, noun	-ist(ic), adjective
(1) Darwin-	the theory of evolution based on Charles Darwin's theories about the gradual development of plant and animal species by natural favouring of individuals with certain characteristics	a supporter of Darwinism	dealing with or relating to Darwinism
(2) aut-	a developmental disorder manifested by communication and language difficulties, problems with social interaction and repetitive patterns in behaviour and interests	a person who suffers from <autism>	suffering from <autism>; relating to autism
(3) universal-	the principle or doctrine (e.g., in religion or politics) that emphasises the precedence of the common good or the whole over the individual or the specific	a person who is a supporter of <universalism>, believing in the precedence of the common good or the whole over the individual or the specific	relating to or characteristic of <universalism>, i.e., the principle that the common good or the whole takes precedence over the individual or the specific
(4) Macchiavell-	a political doctrine which asserts that the State has the right to use any means to achieve its goals; amoral or reckless politics	a person who believes that an individual politician or the State has the right to use any means to achieve their goals	engaging in amoral or ruthless politics and believing that those in power have the right to use any means to achieve their goals

In the present study, we compared the 33 DDO triplets, or 99 headwords, with the corresponding entries in the other dictionaries to see if their definitions could be assigned to one of the four templates. In Appendix 1, all the triplets and their classification are listed for the interested reader.

But first a few words about the data and the methodology. Although the four templates can be seen as lying on a continuum, with one end most adapted to the paper dictionary and the other most to electronic publishing – or, put more neutrally, ranging from fully dependent to fully independent definitions – it does not make sense to quantify the findings beyond counting and ranking: a triplet that follows Template 4 is not twice as adapted to electronic publishing as one following Template 2. For this reason, and due to the limited scope of the material, we content ourselves with describing the trends we can observe in the material by simple counts.

If we turn to look at the results, let us first consider lexical coverage. As the headwords were selected from DDO, it should come as no surprise that not all of them are found in the other dictionaries. This says, of course, little about the lexical coverage of the dictionaries in general; if the original selection had been made from one of the others, the coverage would have looked correspondingly different. But for the sake of comparison, it is worth mentioning how many of the 99 words have an equivalent in the other dictionaries:

DDO: 99

DDNO: 88

NAOB: 90

SO: 71

ODE: 68

Absence of a headword can either mean (a) that no result at all is returned for a queried word, or (b) that a different entry is returned in which the queried word appears. In the first case, the user is likely to be disappointed, whereas the second seems to be used as a deliberate strategy by DDNO and ODE: a search for *shintoist* leads in both dictionaries to the entry *Shinto* where the derived word forms *Shintoism*, *Shintoist* and (in DDNO) *Shintoistic* appear. The dictionary editors may assume that the users are aware where they are and why they arrived in that entry. Especially the ODE has several examples of this strategy: *dualism*, *Fauvism*, *functionalism*, *pietism*, *behaviourism*, *Darwinism*, *multiculturalism*, *social realism*, *Zen buddhism*, *Dada* and *Shinto* are all examples of triplets which have been merged into one headword, and the corresponding person and adjective are only mentioned there as derivatives. One should also be aware that, due to English morphology, what are triplets in the Scandinavian languages are often only doublets in English: there is no independent adjective *secularistic* in ODE; instead the word form *secularist* is used both for the person (as a noun) and as a (preposed) adjective. These factors contribute to explaining the relatively low coverage in ODE.

Table 2: Distribution of triplets/individual definitions across template type¹

Triplets	DDO	DDNO	NAOB	SO	ODE
Template 1	10/28	3/9	6/26	2/18	0/6
Template 2	2/10	0/0	8/26	1/13	0/0
Template 3	4/10	0/0	0/1	0/0	0/0
Template 4	4/13	23/50	1/3	0/4	7/18
Total	20/61	26/59	15/56	3/35	7/24

In Table 2, two numbers are given for each cell: the first shows the number of consistent triplets, i.e., where the same template is used for both derived forms. The second is the overall number of times this template was used, irrespective of the template used for the other derived form. The findings should be read with some caution and can only be used to illustrate tendencies.

Firstly, it should be noted that determining the number of consistent triplets is not straightforward and involves some personal judgment. This is because the templates are generalisations over ideal types. In reality, the definitions may be less ‘pure’ and may use a combination of features from different templates. Consider the definitions of *Shintoist* and *Zionist* in DDO (in our translation):

¹ As only two members of a triplet can be derivations, the maximum number of definitions are 33 times 2 = 66. The difference between the maximal number and the total number listed are absent lemmas and indeterminate cases.

- (1) *Shintoist*: a person who adheres to the Japanese religion Shinto
- (2) *Zionist* (adj.): relating or belonging to the Jewish political movement <Zionism>

(1) is almost like the minimalistic definition of Template 1, and (2) is almost like the minimalistic definition of Template 2 (like Template 1 but with a clickable link). However, to both definitions a little bit of extra explanation has been added, and if one emphasises this aspect, they approximate Template 3: we are told that Shinto is a Japanese religion, and that Zionism is a Jewish political movement. Is this sufficient to say that they can function as independent definitions? We are not sure, and in Table 2 such indeterminate cases were excluded from the count.

This means that a low number of consistent triplets in Table 2 could be due to 1) inconsistent use of templates, e.g., the derived form in the sense ‘person’ follows Template 1 and the derived adjective form follows Template 2, 2) uncertainty about which template to assign (should *Shintoist* in DDO be classified as Template 1 or 3?), or 3) the headword is either absent or has only one entry, in which case consistency becomes irrelevant. If any of these conditions was met, the triplet was left out of the count in Table 2.

With these reservations in mind, it is still safe to say that DDNO is the most consistent in its solutions: out of the 33 triplets, three had only one headword (*creationism*, *secularism*, and *Shinto*), and of the 30 remaining, only three were inconsistent in their choice of template. For the remaining 27, three used the minimalistic Template 1, and as many as 24 followed Template 4, which may lead us to conclude that DDNO prefers Template 4 for this type of definition, and has almost completed its implementation. The three remaining exceptions could be a relic from an earlier phase.

On the surface, ODE is also quite consistent: all consistent triplets follow Template 4. However, the total number of consistent triplets is not very high, only seven, which may be attributed to different factors. We have already mentioned that ODE often has only one entry for the central concept with the derivatives listed there. This is the case for twelve of the triplets, and for another three, there is no headword at all. The remaining eleven are either indeterminate (six) or inconsistent (five). Thus, we must conclude that ODE prefers either the single-headword solution or a solution with Template 4, but bear in mind that a third of the material deviates from this by being either indeterminate or inconsistent.

NAOB is consistent in slightly less than half of the cases and does not seem to have a clear preference for a particular solution, apart from opting out of Template 3. It is important to note, however, that NAOB is also the oldest dictionary and therefore has a significant historical backlog to address. The quickest way to update the rather poor relic of the paper dictionary in Template 1 is to turn the core concept into a hyperlink, which could explain why NAOB seems to give preference to Template 2.

The very low number of consistent triplets in SO probably has more than one explanation. Firstly, SO is the smallest of the included dictionaries (approximately 65,000 entries) and therefore lacks some of the DDO entries. Of the investigated 99 headwords, SO only has 71, which means that the number of incomplete triplets is substantial. Among the remaining headwords, quite a few are inconsistent (nine triplets), typically using both Template 1 and Template 2 within the same triplet. Template 3 is never used, and Template 4 is only used in four individual definitions (and not consistently in any triplet). Finally, two triplets were omitted because of indeterminacy. Overall, SO is the most traditional of the four, with a preference for the minimal solution for morpho-semantic derivations, distributed more or less equally between no link and link (Templates 1 and 2).

For the sake of completeness, it must be noted that characterising DDO in the same way as the other dictionaries is not appropriate: the entries were selected from DDO in such a way that, firstly, we would get complete triplets for all the words and, secondly, that we would get a wide selection of solutions to the current problem. As they were not randomly selected, we cannot know if they are representative of all the words of this type. DDO has more than 500 headwords with *-isme*, so the selected 33 constitute just 6–7% of the total number. Suffice it to say here that since we address the problem, we are aware of it, and it is evident that there are inconsistencies in DDO, both internally in the choice within the individual triplets and more generally in the editorial choice of preferred template.

To sum up the tendencies in the dictionaries examined, DDO may take comfort in the fact that it is not alone in being inconsistent in this area. Even the most consistent of the dictionaries investigated, DDNO, only used the same template in 24 (or approximately 72%) out of the 33 cases under investigation, or out of the 30 that were included in the dictionary (in which case 80%). Another finding is that none of the other dictionaries makes consistent use of Template 3. Even in the individual definitions, Template 3 is only used in eight instances outside of DDO, and they are all instances of indeterminacy. This is slightly surprising to us as we find this solution to be a good one, showing both the morphological relatedness and unfolding the semantic description in an independent definition.

4.2 Investigating Inheritance

As mentioned in Section 3, it is common, at least in paper dictionaries, that secondary or subordinate senses presuppose (inherit) elements from a sense treated earlier in the entry. In order to examine this phenomenon, we selected twenty entries from The Danish Dictionary and compared them to the equivalent entries in the other four dictionaries (see Appendix 2 for a full list). We found the entries by means of specific search patterns identifying subordinate definitions including words like *sådan* ('such'), *denne/dette* ('this') and *disse* ('these'). Such definitions make use of anaphoric reference to other parts of the dictionary entry. In order to find other solutions to the problem, these searches were supplemented by (a) queries for subordinate senses belonging to the domain *food* with hyperlinks in

their definitions and (b) queries for senses belonging to the domains *zoology* and *botany* with words like ‘food’, ‘meat’, ‘domestic animal’, ‘eat’ and ‘taste’ in their definitions. One should be aware that, just as in the previous section, this means that the twenty entries from DDO were not randomly selected. However, while any tendency found for DDO could just be a consequence of the selection procedure, this is not the case for the other dictionaries: their solutions are not dependent on DDO’s description.

We have identified four different templates or rather types of solutions used by the different dictionaries:

Table 3: Types of definitions of subordinate senses (translated from Danish except for Type 4)

Types	Definitions
Type 1 (anaphoric reference)	<i>agate</i> : piece of this mineral, for instance used as a jewel (DDO)
Type 2 (use of the headword with or without a hyperlink)	<i>lamb</i> : meat from a lamb (DDO) <i>Hamburg parsley</i> : root from the plant <Hamburg parsley>, used as a vegetable, for instance in a soup (DDO)
Type 3 (independent definition)	<i>lamb</i> : meat from a young sheep (DDNO)
Type 4 (several meanings merged into one) ²	<i>yak</i> : a large domesticated wild ox with shaggy hair, humped shoulders, and large horns, used in Tibet as a pack animal and for its milk, meat, and hide (ODE)

As suggested above, we note but do not place much importance on the fact that DDO has a strong preference for Type 1 (see Table 4), where definitions for subordinate senses include an anaphoric reference to a previous definition as in the example *agate*. The finding for DDO is trivial since the examined entries were selected exactly to illustrate this phenomenon, and most of the selected words were chosen from this type.

The other four dictionaries as a whole use the anaphoric reference considerably less but adopt other strategies. Two of the dictionaries, DDNO and NAOB, seem to have abandoned the traditional template with an anaphoric expression and do not use it at all, while ODE only does so for a single word. SO is somewhat closer to DDO, using anaphors in a few cases, but rarely employs the other types. Typically, the subsense is either absent, or the sense division handled differently.

DDNO seems to be quite fond of Type 3, independent definitions, a tendency we already saw in Section 4.1. NAOB, on the other hand, makes rather extensive use of Type 2, reference to the headword with or without link. Finally, ODE seems also to prefer mentioning the headword under the subordinate sense (Type 2) but, somewhat surprisingly, never with a link.

² Even if the subordinate senses are not singled out explicitly, the derived senses are at least hinted and may help the user understand examples such as ‘yak meat’ or ‘yak leather’.

Table 4: Distribution of types across inheritance examples

Types	DDO	DDNO	NAOB	SO	ODE
Type 1	14	0	0	4	1
Type 2	2	2	8	1	5
Type 3	0	5	0	1	4
Type 4	1	3	0	1	1
Indeterminate	3	4	5	7	5
No equivalent	--	6	7	6	4
Total	20	20	20	20	20

All in all, the picture is not very clear: the solutions are distributed across a number of possibilities with no clear general preferences. In almost a quarter of the cases, it was not possible to assign the solutions unambiguously to the four categories established here. And in just over a quarter of the cases, the other dictionaries had not registered the relevant sense from DDO, or have chosen a different sense division. As was the case for morpho-semantic similarity, SO is the most traditional of the four, with most examples using the traditional template (Type 1) with anaphoric reference, although the numbers are so small that they cannot be used statistically.

5. Discussion and Conclusions

For both types of definitions, the trends in the examined dictionaries are fairly weak. If we disregard DDO for methodological reasons, the strongest tendency is found for DDNO, which has more or less completely moved away from the paper tradition, with only three traditional triplets remaining out of the thirty found in the dictionary, and nine individual definitions, and having no examples with anaphoric reference in subsenses. ODE tends to follow the same path, with no triplets following the traditional template, six individual definitions, and only a single definition with anaphoric reference. NAOB shares this tendency, but only for inheritance, whereas the picture is reversed for morpho-semantic similarity: here, no less than six triplets consistently follow the traditional template and approximately half of all the definitions for derived lemmas use it. SO has few consistent triplets, but a substantial number of individual definitions follow the minimalist solution, with or without a link. Similarly, for inheritance, the traditional solution predominates, but a substantial number of lemmas and missing indeterminate cases cloud the picture.

For the morpho-semantic triplets, DDO's editors tend to prefer Template 3, which includes mention of the (clickable) *-ism* element and a partially unfolded definition. This is the least preferred solution of the four templates by the other dictionaries.

For inheritance, DDO aligns with most of the other dictionaries: the solution with anaphoric reference (Type 1) should be abandoned, even if it involves violating a traditional *faux pas*: mentioning the headword in another definition of the same word. As long as it happens in a subordinate sense, this price is worth paying. If the

word is made into a hyperlink (as in Type 2), the user has the option of looking it up in a situation where the definition occurs detached from the rest of the entry. In such a situation, Type 3 would probably work better (as it requires no other consultation) but arguably less so in the online dictionary presentation because it involves a bigger processing load for the user. Compare the difference between the two solutions:

Type 2: 1. *lamb*: a young sheep

1.a the meat from a slaughtered <lamb>, used as food

Type 3: 1. *lamb*: a young sheep

1.a the meat from a slaughtered young sheep, used as food

At the time of writing, the editors of the DDO have not yet decided which of the two solutions they prefer.

Finally, it should be noted that what we have shown here, is only the tip of the iceberg. The triplet *-ism*, *-ist*, *-istic* was chosen because it is readily understood internationally, but it is by no means the only type of morpho-semantic similarity. Derivation is, by definition, a productive process, resulting in many other doublets, triplets, etc. through suffixation. We would also like to stress that although we have checked how consistent the dictionaries are in their use of different templates, this does not necessarily constitute a qualitative assessment that a consistent solution is always better. It may well be that different solutions both within and across triplets work fine for the user in the individual case. But we do think that it is important that definitions can be read independently and without the user having to consult other definitions or entries first.

References

- Atkins, B. T. S., & Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford University Press.
- Atkins, B. T. S., Kilgariff, A., & Rundell, M. (2010). Database of Analysed Texts of English (DANTE): the NEID database project. In A. Dykstra, & T. Schoonheim (Eds.), *Proceedings of the 14th EURALEX International Congress* (pp. 549–556). Fryske academy.
- DDO = *Den Danske Ordbog*, The Danish Dictionary, Society for Danish Language and Literature. Retrieved May 31, 2024, from ordnet.dk/ddo
- DDNO = *Den Danske Netordbog*, The Danish Internet Dictionary, Center for Leksikografi. Retrieved May 31, 2024, from ordbogen.com
- Lonke, D., Kernerman, I., & Dzhuranyuk, V. (2022). Lexical Data API. In A. Klosa-Kückelhaus, S. Engelberg, C. Möhrs, & P. Storjohann (Eds.), *Proceedings of the XX EURALEX International Congress* (pp. 401–408). IDS-Verlag.

Marello, C. (1988). Anaphora and deixis as cohesive devices in lexicography. In M. Snell-Hornby (Ed.), *ZüriLEX '86 Proceedings: Papers read at the EURALEX International Congress, University of Zürich, 9–14 September 1986* (pp. 117–123). Francke Verlag.

NAOB = *Det Norske Akademis Ordbok*, Dictionary of the Norwegian Academy, Det Norske Akademi for Språk og Litteratur. Retrieved May 31, 2024, from naob.no

ODE = *Oxford Dictionary of English* (3 ed.), Oxford University Press. Retrieved May 31, 2024, from www.oxfordreference.com

SO = *Svensk ordbok*, Swedish Dictionary, Svenska Akademien. Retrieved May 31, 2024, from svenska.se

Svensén, B. (2009). *A Handbook of Lexicography. The Theory and Practice of Dictionary-Making*. Cambridge University Press.

Sørensen, N., Nimb, S., & Pedersen, B. S. (2023). How do We Treat Systematic Polysemy in Wordnets and Similar Resources? – Using Human Intuition and Contextualized Embeddings as Guidance. In G. Rigau, F. Bond, & A. Rademaker (Eds.), *Proceedings of the 12th Global Wordnet Conference, University of the Basque Country, Donostia – San Sebastian, Basque Country. Global Wordnet Association* (pp. 117–126). Global Wordnet Association.

Contact information

Lars Trap-Jensen

Society for Danish Language and Literature
ltj@dsl.dk

Henrik Lorentzen

Society for Danish Language and Literature
hl@dsl.dk

Appendix 1) Triplets and Templates

	Lemma	DDO	DDNO	NAOB	SO	ODE
1a	absurdisme	yes	yes	yes	yes	yes
1b	absurdist	yes (1)	yes (4)	yes (1)	no	yes (0)
1c	absurdistisk	yes (1)	yes (4)	yes (1)	no	yes (4)
2a	aktivisme	yes	yes	yes	yes	yes
2b	aktivist	yes (4)	yes (4)	yes (0)	yes (4)	yes (4)
2c	aktivistisk	yes (1)	yes (4)	yes (2)	no	yes (4)
3a	alkymi	yes	yes	yes	yes	yes
3b	alkymist	yes (2)	yes (4)	yes (1)	yes (1)	yes (1)
3c	alkymistisk	yes (2)	yes (4)	yes (2)	no	no
4a	autisme	yes	yes	yes	yes	yes
4b	autist	yes (2)	yes (1)	yes (2)	yes (1)	no
4c	autistisk	yes (2)	yes (1)	yes (2)	yes (1)	yes (1)
5a	behaviorisme	yes	yes	yes	yes	yes
5b	behaviorist	yes (1)	yes (4)	yes (0)	yes (1)	no (but mentioned in (a))
5c	behavioristisk	yes (1)	yes (4)	yes (1)	no	no (but mentioned in (a))
6a	calvinisme	yes	yes	yes	yes	yes
6b	calvinist	yes (1)	yes (4)	yes (2)	yes (2)	yes (0)
6c	calvinistisk	yes (1)	yes (4)	yes (2)	yes (1)	yes (0)
7a	dadaisme	yes	yes	yes	yes	yes
7b	dadaist	yes (1)	yes (4)	yes (2)	yes (2)	no (but mentioned in (a))
7c	dadaistisk	yes (1)	yes (4)	yes (2)	yes (1)	no (but mentioned in (a))
8a	darwinisme	yes	yes	yes	yes	yes
8b	darwinist	yes (1)	yes (1)	yes (2)	yes (2)	no (but mentioned in (a))
8c	darwinistisk	yes (1)	yes (1)	yes (1)	yes (2)	no (but mentioned in (a))
9a	dualisme	yes	yes	yes	yes	yes
9b	dualist	yes (2)	yes (4)	yes (1)	yes (2)	no (but mentioned in (a))
9c	dualistisk	yes (1)	yes (4)	yes (2)	yes (1)	no (but mentioned in (a))
10a	fauvisme	yes	yes	yes	yes	yes
10b	fauvist	yes (3)	yes (1)	yes (2)	no	no (but mentioned in (a))
10c	fauvistisk	yes (3)	yes (4)	yes (2)	no	no (but mentioned in (a))
11a	funktionalisme	yes	yes	yes	yes	yes
11b	funktionalist	yes (3)	yes (4)	yes (1)	yes (1)	no (but mentioned in (a))
11c	funktionalistisk	yes (3)	yes (4)	yes (1)	yes (2)	no (but mentioned in (a))
12a	kolonialisme	yes	yes	yes	yes	yes
12b	kolonialist	yes (1)	yes (4)	yes (2)	no	yes (4)
12c	kolonialistisk	yes (1)	yes (4)	yes (2)	no	yes (4)
13a	konkretisme	yes	yes	yes	yes	no
13b	konkretist	yes (1)	yes (4)	yes (2)	no	no
13c	konkretistisk	yes (1)	yes (4)	yes (2)	no	no

14a	kreationisme	yes	yes	no	yes	yes
14b	kreationist	yes (1)	no (but mentioned in (a))	no	yes (1)	yes (4)
14c	kreationistisk	yes (1)	no (but mentioned in (a))	no	yes (1)	yes (4)
15a	machiavellisme	yes	yes	no	yes	no
15b	machiavellist	yes (4)	yes (4)	no	no	no
15c	machiavellistisk	yes (4)	yes (4)	no	yes (4)	no
16a	maoisme	yes	yes	yes	yes	yes
16b	maoist	yes (3)	yes (1)	yes (1)	no	yes (0)
16c	maoistisk	yes (2)	yes (1)	yes (1)	no	yes (0)
17a	modernisme	yes	yes	yes	yes	yes
17b	modernist	yes (0)	yes (4)	yes (2)	yes (0)	yes (0)
17c	modernistisk	yes (0)	yes (4)	yes (1)	yes (0)	yes (0)
18a	multikulturalisme	yes	yes	yes	yes	yes
18b	multikulturalist	yes (4)	yes (4)	yes (1)	no	no (but mentioned in (a))
18c	multikulturalistisk	yes (4)	yes (4)	yes (2)	no	no (but mentioned in (a))
19a	nynazisme	yes	yes	yes	yes	no (but mentioned in (a))
19b	nynazist	yes (1)	yes (4)	yes (2)	yes (2)	yes
19c	nynazistisk	yes (1)	yes (4)	yes (2)	no	yes (1)
20a	nihilisme	yes	yes	yes	yes	yes
20b	nihilist	yes (1)	yes (1)	yes (1)	yes (2)	yes (4)
20c	nihilistisk	yes (1)	yes (4)	yes (2)	yes (1)	yes (4)
21a	pietisme	yes	yes	yes	yes	yes
21b	pietist	yes (4)	yes (4)	yes (2)	yes (1)	no (but mentioned in (a))
21c	pietistisk	yes (0)	yes (4)	yes (2)	yes (2)	no (but mentioned in (a))
22a	pluralisme	yes	yes	yes	yes	yes
22b	pluralist	yes (2)	yes (4)	yes (1)	no	yes (4)
22c	pluralistisk	yes (1)	yes (4)	yes (1)	yes (4)	yes (4)
23a	relativisme	yes	yes	yes	yes	yes
23b	relativist	yes (4)	yes (4)	yes (1)	yes (1)	no (but mentioned in (a))
23c	relativistisk	yes (4)	yes (4)	yes (0)	yes (2)	yes (1)
24a	royalisme	yes	yes	yes	yes	no (but mentioned in (a))
24b	royalist	yes (4)	yes (4)	yes (4)	yes (4)	yes
24c	royalistisk	yes (4)	yes (4)	yes (4)	yes (0)	yes (4)
25a	salafisme	yes	yes	yes	yes	yes (as Salafi)
25b	salafist	yes (3)	yes (4)	yes (1)	yes (2)	no (but mentioned in (a))
25c	salafistisk	yes (3)	no (but mentioned in (a))	yes (1)	no	no (but mentioned in (a))
26a	sekularisme	yes	yes	yes	no	yes
26b	sekularist	yes (4)	no	no	no	yes (4)
26c	sekularistisk	yes (2)	no	no	no	yes (4)
27a	shintoisme	yes	yes	yes	yes	yes (as Shinto)
27b	shintoist	yes (0)	no (but mentioned in (a))	yes (1)	no	no (but mentioned in (a))
27c	shintoistisk	yes (1)	no (but mentioned in (a))	yes (2)	no	no

28a	socialrealisme	yes	yes	yes	yes	yes
28b	socialrealist	yes (4)	yes (4)	no	no	no
28c	socialrealistisk	yes (1)	yes (4)	yes (1)	yes (1)	no
29a	stalinisme	yes	yes	yes	yes	yes
29b	stalinist	yes (3)	yes (4)	yes (1)	yes (2)	yes (0)
29c	stalinistisk	yes (1)	yes (4)	yes (1)	yes (1)	yes (0)
30a	surrealisme	yes	yes	yes	yes	yes
30b	surrealist	yes (2)	yes (4)	yes (0)	no	yes (4)
30c	surrealistisk	yes (1)	yes (4)	yes (1)	yes (1)	yes (4)
31a	universalisme	yes	yes	yes	yes	yes
31b	universalist	yes (3)	yes (4)	yes (2)	yes (1)	yes (4)
31c	universalistisk	yes (3)	yes (4)	yes (1)	yes (2)	yes (1)
32a	zenbuddhisme	yes	yes	yes	yes	yes
32b	zenbuddhist	yes (1)	yes (4)	yes (2)	no	no
32c	zenbuddhistisk	yes (2)	yes (1)	yes (1)	no	no
33a	zionisme	yes	yes	yes	yes	yes
33b	zionist	yes (4)	yes (4)	yes (3)	yes (1)	yes (4)
33c	zionistisk	yes (0)	yes (4)	yes (4)	no	yes (1)

Templates:

1. indeterminate
2. minimalist definition: a person who supports x
3. minimalist definition with link: a person who supports <x>minimalist definition with a link and short explanation of the link: a person who supports <x>, the belief that y
4. fully expanded definition: a person who believes that life is meaningless and rejects all religious and moral principles

Appendix 2) Inheritance and Types

	Lemma	Inherited sense	DDO	DDNO	NAOB	SO	ODE
1	ablativ	wordform in this case	yes (1)	no equivalent	no equivalent	no equivalent	yes (3)
2	afgrøde	yield from this cultivation	yes (1)	no equivalent	yes (2)	yes (3)	yes (3)
3	afsender	this person's name and address	yes (1)	yes (2)	no equivalent	no equivalent	no equivalent
4	agat	piece of this mineral	yes (1)	yes (4)	yes (2)	indeterminate	indeterminate
5	ambitiøs	indicating that someone has such a wish	yes (1)	no equivalent	no equivalent	indeterminate	yes (3)
6	gulerod	root from this plant, used as vegetable	yes (1)	indeterminate	yes (2)	yes (1)	yes (1)
7	kalv	a) meat from a calf; b) skin from a calf	indeterminate	indeterminate	indeterminate	indeterminate	indeterminate
8	lam	meat from a lamb	yes (2)	yes (3)	indeterminate	indeterminate	yes (2)
9	lama	textile made from llama wool	indeterminate	indeterminate	indeterminate	indeterminate	indeterminate
10	persillerod	root from the plant Hamburg parsley	yes (2)	yes (4)	indeterminate	no equivalent	no equivalent
11	pyramide	building formed like such a monument	yes (1)	yes (3)	yes (2)	yes (1)	yes (2)
12	pædagogik	subject where this domain is taught	yes (1)	yes (3)	no equivalent	indeterminate	yes (3)
13	raclette	cheese used for this type of food	yes (1)	yes (3)	yes (2)	no equivalent	no equivalent
14	radise	root from this plant	yes (1)	no equivalent	indeterminate	indeterminate	yes (2)
15	rastafari	follower of this movement	yes (1)	yes (2)	yes (2)	yes (2)	yes (2)
16	shiraz	wine made from this grape	yes (1)	indeterminate	no equivalent	no equivalent	indeterminate
17	skole	pupils and employees of such an institution	yes (1)	yes (3)	indeterminate	yes (1)	yes (2)
18	sortkommen	seeds from the plant Nigella	indeterminate	no equivalent	no equivalent	no equivalent	no equivalent
19	torsk	this fish used as food	yes (1)	no equivalent	yes (2)	yes (1)	indeterminate
20	yak	domestic ox with long hair kept as a domestic animal	yes (4)	yes (4)	no equivalent	yes (4)	yes (4)

Types:

1. anaphoric reference to a previous sense
2. use of the headword with or without a hyperlink
3. independent definition
4. several meanings merged into one

Ene Vainik, Geda Paulsen, Heete Sahkai, Jelena Kallas, Arvi Tavast, and Kristina Koppel

FROM A DICTIONARY TO A CONSTRUCTICON

Putting the Basics on the Map

Abstract This study discusses the possibilities of expanding the scope of the largest Estonian dictionary – the EKI Combined Dictionary – with various types of constructional information. Designing a representation of constructions essentially means building a constructicon. The study starts with a short overview of existing constructicons and the main challenges their creators have faced so far. We address these issues from the point of view of data model reorganisation and database restructuring. Extending the lexicographic resource with constructicographic information is twofold: the existing constructional information must be migrated into a new model and then complemented with additional constructions extracted from a corpus.

Keywords electronic lexicography; grammatical constructions; constructicography; Estonian

1. A New Emerging Field in Electronic Lexicography

As electronic lexicography continues to evolve, a new subfield called constructicography has emerged (Lyngfelt, 2018, p. 1). This field focuses on compiling a resource known as a “constructicon”, which parallels dictionaries by presenting complex grammatical constructions as pairs of linguistic forms and meanings. Unlike traditional lexicography, constructicography is theory-driven and rooted in the construction-based view of language (e.g., Croft, 2001; Goldberg, 2003; Hoffmann & Trousdale, 2013), asserting that there is a continuum of linguistic units instead of a lexicon vs. grammar dichotomy and that networks of constructions are fundamental to language (Diessel, 2023). Constructicons aim to not only present the form and meaning of their units but also their interrelations. The use of relational databases facilitates this endeavour. Though approximately ten initiatives worldwide are underway to create such resources for various languages, there is currently no established tradition in this field (Borin & Lyngfelt, in press).

This paper aims to contribute to constructicographic theory and practice by analysing the decisions and challenges of current initiatives (i.e., mapping the basics; see Section 2). In Section 3, we formulate our initial plan to enrich the largest Estonian dictionary, the EKI Combined Dictionary (CombiDic)¹ (Koppel et al., 2019), with information about constructions and present the main points of restructuring the data model to meet the needs of a constructicographic resource.

¹ Accessible via the language portal sõnaveeb.ee

2. Mapping the Basics

This section provides an overview of current constructicographic initiatives (Fillmore et al., 2012; Janda et al., 2020; Lyngfelt, Bäckström, et al., 2018; Ohara, 2013; Perek & Patten, 2019; Sass, 2023; Torrent et al., 2014; Ziem et al., 2019), focusing on their target groups, size, types of relations, database types, and connected resources (see Appendix 1). We also list the main challenges reported by contributors and those we have identified.

2.1 The Current Practices

The analysis of current practices showed that most of the initiatives have grown out of ongoing work on a FrameNet database and are linked to it. The reported coverage of constructions varies from 73 to approx. 13 000 entries. The target users can be language experts, L2 learners, language technology or some combination of those. The primary focus varies from semi-schematic and/or idiosyncratic units to abstract valency patterns (argument structure constructions) and, to a somewhat lesser degree, to idiomatic units residing on the “lexical end” of the grammar-lexicon continuum. An essential part of constructicon building is presenting the relations of inheritance (predicted by the theory, e.g., Ziem et al., 2023) and, to a lesser degree, part-whole relationships (e.g., Sass, 2023).

2.2 The Challenges

The process of creating constructicons is challenging and time-consuming, involving various hurdles:

1. Defining what qualifies as a construction for inclusion. While all meaning-form pairs in a language are theoretically constructions, it's crucial to identify suitable constructicon units. This relates to determining the complete inventory of constructions to be included.
2. Presenting constructions at different degrees of schematicity. The data model and description format must be suitable for describing units at various points along the lexicon-grammar continuum, from idioms and phraseological units to intermediate forms and purely schematic grammatical constructions (Ziem et al., 2023).
3. What should a constructicon entry look like? Language learners require simplified descriptions, meta-language and only typical usage examples, while experts seek detailed information and annotated corpus sentences. Ensuring machine readability is another aspect related to the granularity of description. Therefore, a constructicon entry must be adequate, concise, user-friendly, and at the same time formalised (Lyngfelt, Borin, et al., 2018).
4. Naming constructions and organising constructicon units. Since alphabetical order may not suffice, determining organisation criteria, such as taxonomies or other typological systems, is necessary. Should the presentation of

categories mimic that of grammar? Should entries be allowed to appear in multiple types?

5. Representing relationships between constructions. Different constructicons adopt varied approaches, reflecting the dataset's purpose: what is suitable for a pedagogically oriented constructicon may not be suitable for an NLP-oriented solution. The task of presenting slot-filler relations is a challenge because the relations must be established first.
6. How is it possible to make information about constructions accessible to users with different linguistic backgrounds? While lexical units can be given translational equivalents, how should a user find the most natural way to express desired meanings in the target language from the constructicon? (see e.g., Lyngfelt et al., 2022).

Borin & Lyngfelt (in press) emphasise that in addition to structural elements derived from theory, it is important to consider whether the constructicon is associated with another lexicographic dataset (such as FrameNet) and what is the overall purpose of the constructicon. The structure must be flexible and adaptable so that new (and possibly different types of) constructions can be added to the collection.

3. Towards a Data Model

The Estonian project primarily targets L2 learners and teachers, along with L1 speakers, researchers, and NLP applications. Due to the lack of a FrameNet for Estonian and the need for a comprehensive lexical and grammatical resource, we are developing the Estonian constructicon as an extension of CombiDic (Koppel et al., 2019). CombiDic is the most extensive and up-to-date Estonian lexicographic data collection, containing a full lexical inventory, morphological paradigms, definitions, syntactic information, semantic types and CEFR proficiency level markers. It has been compiled using the Ekilex Dictionary Writing System (Tavast et al., 2018) on a PostgreSQL database.

3.1 The Data Model

To describe the representation of constructions, we first need to outline the Ekilex data model², which is based on a many-to-many relationship between word and meaning. In database terms, such relations are implemented using junction tables. The junction table between word and meaning is called a lexeme, and the data it contains can be described as “this word in this meaning as defined by this dictionary”.

The data model already includes the following syntactic data elements: i) government patterns: currently a plain text field within the lexeme; ii) compounds and derivatives: relations between words, e.g., compounds containing information about their components; iii) collocations: until recently separate units closely duplicating the representation of words.

² <https://github.com/keeleinstituut/ekilex/wiki/Andmemudel>

The first stage of extending Ekilex to accommodate a constructicon involves redesigning the data model for collocations and migrating existing collocations to this new model, along with a more robust representation of government. Eventually, compounds and derivatives will also be unified into the same representation.

As discussed above, a consensus about what exactly is a constructicon is yet to emerge. Our current approach is to start with the corpus, i.e., actualisations of the constructions, and move towards abstraction. As a result, instead of being standalone, i.e., unrelated to a dictionary (e.g., Janda et al., 2020), linked to a dictionary (e.g., Fillmore et al., 2012; Lyngfelt, Bäckström, et al., 2018; Ohara, 2013; Perek & Patten, 2019; Torrent et al., 2014; Ziem et al., 2019), or a replacement for a dictionary (Sass, 2024), we conceptualise the constructicon as something that a dictionary can grow into when expanded by syntactic information.

At the heart of the redesign is the idea of expanding the concept of the headword to include units of language of any length, notably including units that have so far been presented separately as collocations or compounds. Although this increases the number of headwords, it is not a big change conceptually, as many headwords have always been multi-word units. Each expanded headword has parameters controlling its presentation: it can be displayed as a separate headword with its own entry and/or as a collocation and/or compound within the entry of one or more of its components. Just like lexemes, these relations are meaningful data structures on their own rather than simple links between the whole and its parts. They are the data elements with which we implement abstraction from individual realisations of a construction towards its schematic representation. Relations can be tagged using multiple, mutually independent tagging systems, aiming to represent existing morphological and syntactic hierarchies and support ongoing construction grammar research.

A major challenge is the rich morphology of Estonian, including pervasive paradigmatic homonymy, the disambiguation of which has not yet been reliably solved. For sustainability and scalability, we aim for fully structured data in Ekilex, connecting each MWE component to its form, not just the lexeme. This has required manual disambiguation and minor changes to the morphology section of the data model.

3.2 Meeting the Challenges

In the following, we summarise our approach to address the challenges outlined in Section 2.2.

1. In terms of determining what qualifies as a construction to be included, we will start by reorganising and generalising the existing syntactic information in the CombiDic, specifically government patterns and collocational patterns.
2. To represent constructions at different degrees of schematicity, our current approach is to detect phrase structure and argument structure constructions from the morphologically and dependency syntactically annotated corpus and move towards abstraction, thereby achieving a hierarchical representation.

3. The structure of constructicon entries should cater to the main user groups: concise and simplified for language learners, detailed for experts, and formalised for NLP. This will be achieved by designing a comprehensive entry format with customised interfaces for different users.
4. Regarding the challenge of naming and organising the constructions, we plan to establish a typology of constructions and tag the entries with a set of labels (e.g., parallel names for different user groups, as well as taxonomic/typological and semantic affiliation). Multiple tags per construction will be allowed to facilitate recall by different criteria.
5. Hierarchical relations between entries will be crucial in our data model. The dictionary-based approach makes it possible to establish relations with lexemes as slot-fillers in semi-schematic and schematic constructions. Existing word class and semantic type labels can represent these filler-slot relations.
6. To make constructional information accessible to different user groups, we should integrate it with the general search, potentially including free text search (cf. Sass, 2023), drop-down menus for semantic categories (cf. Janda et al., 2020), a filterable list of all the constructions, and access via a layer of syntactic information in lexical entries.

4. Conclusion

Each constructicon project is unique, but most face similar challenges. This article summarises current constructicographic practices and compares them to our approach for extending a lexicographic resource with constructicographic information. Challenges include defining and naming constructions, and how to represent and display them. The data model must be flexible enough to cover the entire lexicon-grammar continuum. Solutions to these challenges inform the design for representing constructions in the EKI Combined Dictionary and its Ekilex database. Our goal is to redefine existing lexical and grammatical information in Ekilex and enrich it with (semi-)schematic constructions, while addressing the additional complexity of Estonian morphology.

References

- Borin, L., & Lyngfelt, B. (in press). Framenets and constructiCons. In *The Cambridge Handbook of Construction Grammar*. Cambridge University Press. https://www.academia.edu/95301779/Framenets_and_constructiCons
- Croft, W. (2001). *Radical Construction Grammar: Syntactic Theory in Typological Perspective* (1st ed.). Oxford University Press/Oxford. <https://doi.org/10.1093/acprof:oso/9780198299554.001.0001>
- Diessel, H. (2023). The Constructicon: Taxonomies and Networks. In *Elements in Construction Grammar*. <https://doi.org/10.1017/9781009327848>

Fillmore, C. J., Lee-Goldman, R., & Rhomieux, R. (2012). *The FrameNet Constructicon*. <https://www1.icsi.berkeley.edu/pubs/ai/framenetconstructicon11.pdf>

Goldberg, A. E. (2003). Constructions: A new theoretical approach to language. *Trends in Cognitive Sciences*, 7(5), 219–224. [https://doi.org/10.1016/S1364-6613\(03\)00080-9](https://doi.org/10.1016/S1364-6613(03)00080-9)

Hoffmann, T., & Trousdale, G. (Eds.). (2013). *The Oxford handbook of construction grammar*. Oxford University Press.

Janda, L. A., Endresen, A., Zhukova, V., Mordashova, D., & Rakhilina, E. (2020). How to build a constructicon in five years. The Russian example. *Belgian Journal of Linguistics*, 34(1), 161–173. <https://doi.org/10.1075/bjl.00043.jan>

Koppel, K., Tavast, A., Langemets, M., & Kallas, J. (2019). Aggregating dictionaries into the language portal Sõnaveeb: Issues with and without a solution. *Proceedings of the eLex 2019 Conference. 1–3 October 2019, Sintra, Portugal. Brno: Lexical Computing CZ, s.r.o., 434–452*. https://elex.link/elex2019/wp-content/uploads/2019/09/eLex_2019_24.pdf

Lyngfelt, B. (2018). Introduction: Constructicons and constructicography. In *Constructicography: Constructicon development across languages* (pp. 1–18). John Benjamins Publishing Company. <https://doi.org/10.1075/cal.22.01lyn>

Lyngfelt, B., Bäckström, L., Borin, L., Ehrlemark, A., & Rydstedt, R. (2018). Constructicography at work. Theory meets practice in the Swedish constructicon. In *Constructicography: Constructicon development across languages* (pp. 41–106). John Benjamins Publishing Company. <https://doi.org/10.1075/cal.22.01lyn>

Lyngfelt, B., Borin, L., Ohara, K., & Torrent, T. T. (Eds.). (2018). *Constructicography: Constructicon development across languages* (Vol. 22). John Benjamins Publishing Company. <https://doi.org/10.1075/cal.22>

Lyngfelt, B., Torrent, T. T., Eli, E. da S. M., & Bäckström, L. (2022). Comparative Concepts as a resource for a multilingual constructicography. In *Valency and constructions: Perspectives on combining words*. Meijerbergs institut för svensk etymologisk forskning.

Ohara, K. (2013). Toward Constructicon Building for Japanese in Japanese FrameNet. *Veredas - Revista de Estudos Linguísticos*, 17(1), 11–27.

Perek, F., & Patten, A. L. (2019). Towards an English Constructicon using patterns and frames. *International Journal of Corpus Linguistics*, 24(3), 354–384. <https://doi.org/10.1075/ijcl.00016.per>

Sass, B. (2023). From a dictionary towards the Hungarian Constructicon. *Electronic Lexicography in the 21st Century (eLex 2023): Invisible Lexicography. Proceedings of the eLex 2023 Conference. Brno, 27–29 June 2023. Brno: Lexical Computing CZ s.r.o.* <https://elex.link/elex2023/wp-content/uploads/105.pdf>

Sass, B. (2024). We need everything: the “ultimate lexical resource” approach to build a constructicon. Book of Abstracts of EAAL 21st Annual Conference, April 18.-19, 2024, Tallinn, Estonia. <https://www.rakenduslingvistika.ee/wp-content/uploads/2024/04/2024-teesid.pdf>

Tavast, A., Langemets, M., Kallas, J., & Koppel, K. (2018). Unified Data Modelling for Presenting Lexical Data: The Case of EKILEX. In J. Čibej, V. Gorjanc, I. Kosem, & S. Krek (Eds.), *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts* (pp. 749–761). Znanstvena založba Filozofske fakultete Univerze v Ljubljani.

Torrent, T. T., Lage, L. M., Sampaio, T. F., Tavares, T. da S., & Matos, E. E. da S. (2014). Revisiting border conflicts between FrameNet and Construction Grammar: Annotation policies for the Brazilian Portuguese Constructicon. *Constructions and Frames*, 6(1), 34–51. <https://doi.org/10.1075/cf.6.1.03tor>

Ziem, A., Böbel, N., & Willich, A. (2023). What's in the constructicon? Relating constructional forms and constructional meanings on the full range of the lexicon-grammar continuum. In *Book of abstracts of 16th International Cognitive Linguistics Conference*. https://iclc16.github.io/abstracts/ICLC16_BoA.pdf

Ziem, A., Flick, J., & Sandkühler, P. (2019). The German Constructicon Project: Framework, methodology, resources. *Lexicographica*, 35(2019), 15–40. <https://doi.org/10.1515/lex-2019-0003>

Acknowledgements

This work was supported by the Estonian Research Council grant (PRG 1978).

Contact information

Ene Vainik

Institute for the Estonian Language
ene.vainik@eki.ee

Geda Paulsen

Institute for the Estonian Language, Uppsala University
geda.paulsen@eki.ee, geda.paulsen@moderna.uu.se

Heete Sahkai

Institute for the Estonian Language
heete.sahkai@eki.ee

Jelena Kallas

Institute for the Estonian Language
jelena.kallas@eki.ee

Arvi Tavast

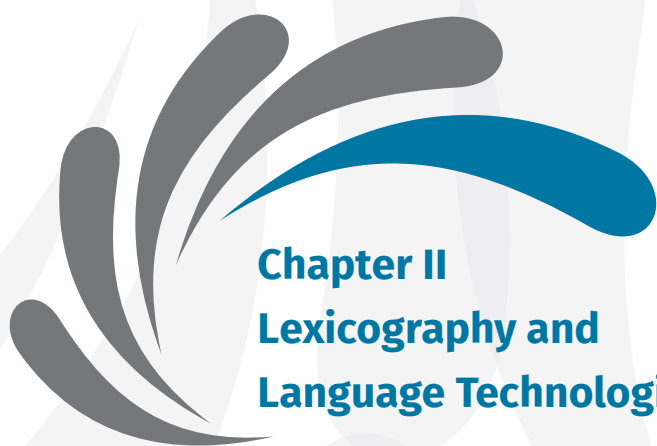
Institute for the Estonian Language
arvi.tavast@eki.ee

Kristina Koppel

Institute for the Estonian Language
kristina.koppel@eki.ee

Appendix 1: Overview of the current initiatives

Language	Coverage	Target users	Focus	Relations	Related resources	Database format	Reference
English	73	Experts	The frame-based description of valency restrictions of lexicographic units	inheritance	Berkeley FrameNet for English	relational	Fillmore et al., 2012
Japanese	ca 50	Experts	The frame-based description of valency restrictions of lexicographic units	inheritance	Japanese FrameNet	relational	Ohara, 2018
Swedish	394	Learners Experts Language technology L2 Teachers	Semischematic constructions; the aim is to cover grammar as whole	inheritance	Swedish FrameNet; SALDO; Karp, and Korp, applications: pedagogical, LT, lexicographic; crosslingual constructicon	relational	Lyngfelt, Bäckström, et al., 2018
Brasilian Portuguese	220	Experts Language technology	The frame-based description of valency restrictions of lexicographic units	(multiple) inheritance	Part of FrameNet for Brazilian Portuguese; the data model has been enhanced	relational	Torrent et al., 2014
Russian	3600	Learners Experts Language technology	Semischematic constructions; the aim is to cover grammar as a whole (from morphemes to the discourse units)		Deduced from reference grammars, corpora and pedagogical materials for teachers and students	relational	Janda et al., 2018
German	225	Experts	Semischematic and semi-idiomatic constructions; the aim is to cover grammar as a whole	various horizontal and vertical	Partly linked to German frame-based dictionary on web	relational	Ziem et al., 2019
English	101	General public Learners	Patterns of extending the relational words (verbs, nouns, adjectives): the phraseology will be treated separately	inheritance	Grammar patterns from COBUILD and semantics from FrameNet	XML	Perek & Patten, 2019
Hungarian	13000	Learners	Morphemes, words, and continuous multiword expressions (without slots)	part-whole	The general dictionary will be turned into constructicon by splitting the existing complex unit and presenting them upon request	XML	Sass, 2023



Chapter II
Lexicography and
Language Technologies



Igor Boguslavsky

LEXICAL RESOURCES FOR A SEMANTIC PARSER OF RUSSIAN Argument Structure of Ordinal Adjectives

Abstract The SemETAP semantic model is a part of a more general ETAP linguistic processor aiming at analyzing and generating NL texts. The task of SemETAP is building two kinds of semantic structures – Basic SemS, which capture the core meaning of the sentence, and Enhanced SemS, which contain diverse inferences drawn from Basic SemS. SemETAP is supported by two main lexical resources – a combinatory dictionary of Russian and an ontology. One important requirement for the SemS is that it should explicitly represent all semantic arguments of the predicates of the sentence, expressed by all kinds of words – verbs, nouns, adjectives or adverbs. We discuss the argument structure of ordinal adjectives (*first, second, ..., last, next*), which has been largely neglected in the literature on valency and arguments. Several semantic slots are introduced for ordinal adjectives: *hasObject*, *hasObject2*, *belongsTo*, *hasNumber*, *hasStartingPoint*, *hasTerminalPoint*, *orderedBy*. Our analysis reveals interesting features in the behavior of the arguments of ordinal adjectives.

Keywords semantic structure; combinatorial dictionary; ontology; argument structure; ordinal adjectives

1. Semantic Parser SemETAP

This work is part of the development of lexical resources elaborated for the SemETAP functional semantic model (Boguslavsky, Frolova et al., 2018; Boguslavsky, Frolova et al. 2019; Boguslavsky, Dikonov et al., 2020). Its immediate task is building semantic representations of Russian texts. Semantic structures (SemS) constructed by SemETAP aim to represent the meaning of the sentence in an explicit and transparent way, while abstracting away from the lexico-syntactic variation of the NL. The model has several salient features.

- SemETAP aims to extract a large number of inferences from a text. We assume that the more inferences we can extract from a text, the more complete and deeper understanding of the text we achieve.
- The model distinguishes between two levels of semantic representation of a sentence: The Basic Semantic Structure, which captures the core meaning of the sentence, and the Enhanced Semantic Structure, which enriches the Basic Structure with a large number of implications. The extension of the Basic Structure is carried out through two main sources: decomposition of concept meanings and common-sense axioms.

- There are two types of implications: strict implicatures and plausible expectations. The latter are particularly important for understanding coherent text and recovering implicit elements of meaning.
- The model is knowledge-based, which means that it relies on explicit knowledge provided by an expert. There are two main lexical resources – a combinatory dictionary of Russian and an ontology, which can be thought of as a semantic dictionary. Both types of semantic structures – the Basic and the Enhanced ones – are built of the elements of the ontology.
- The model functions as a component of a larger system of analysis and generation of texts known as ETAP, which has also a morphological and a syntactic module. In this way, a complete sentence processing cycle is carried out, from the text in orthographic notation to the Enhanced Semantic Structure.
- Basic semantic structures built according to this model constitute a novel semantic corpus (SemOntoCor) that contains more than 2000 sentences at the time of writing and keeps growing (Boguslavsky, Dikonov et al., 2023).

Here is an example to illustrate what SemETAP can do. Sentence (1) is transformed into the Basic SemS shown in Figure 1. (For clarity, green boxes are used to show the correspondence between the words of the sentence and the fragments of the structure).

- (1) *Роналду не спас матч.*
 ‘Ronaldo did not save the match’

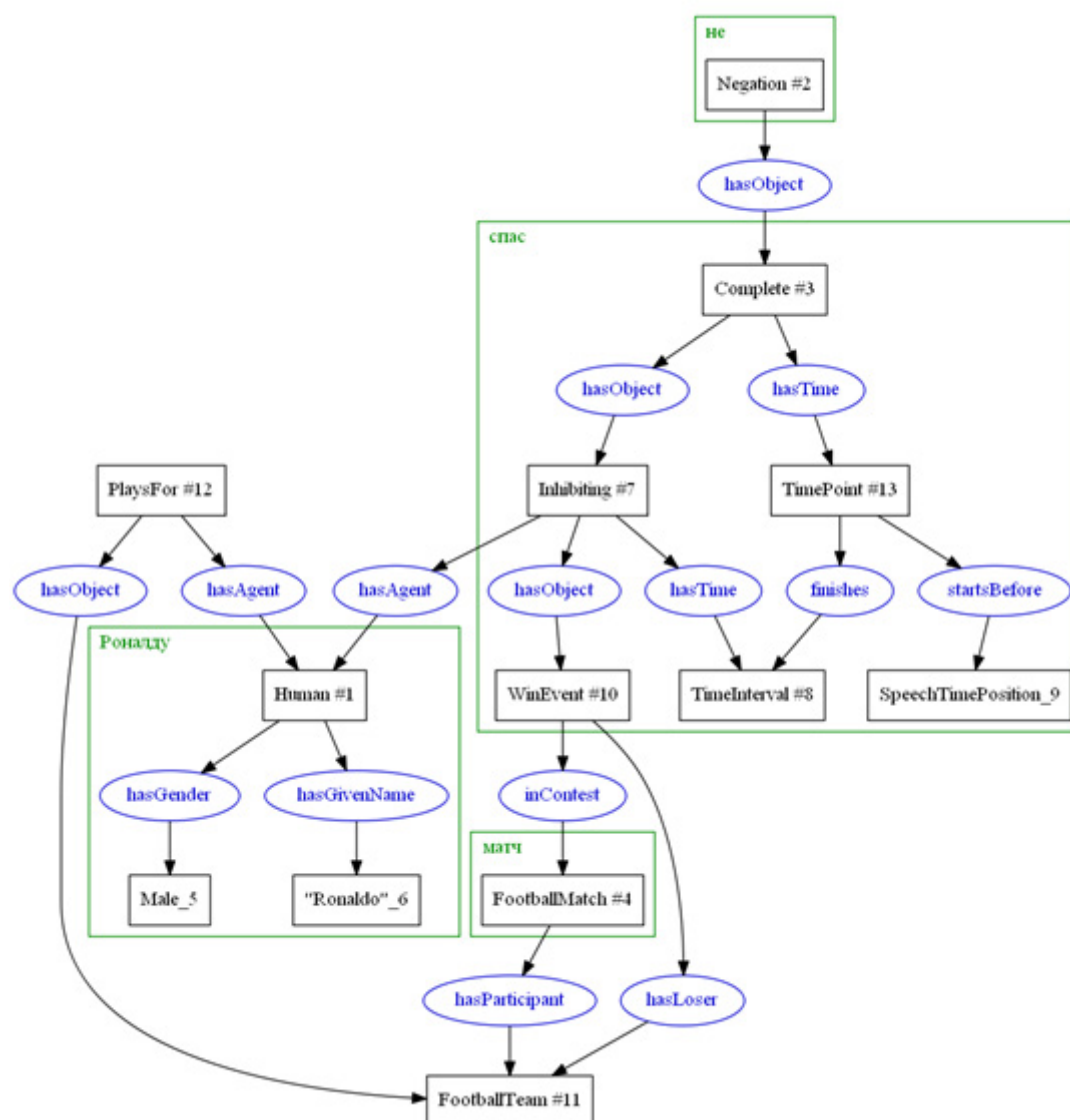


Fig. 1: The Basic semantic structure of the sentence *Роналду не спас матч* 'Ronaldo did not save the match'.

This structure can be “read” as follows: “human #1, whose name is Ronaldo, who is a male and plays for FootballTeam #11, did not inhibit defeat of FootballTeam #11 in a football match in which it participated. All this happened in the past (= before the time of speech)”.

To demonstrate the SemETAP ability to draw inferences, it is convenient to make use of the question-answering mode of the system. In Figure 2, one can see the result of processing sentence (1).

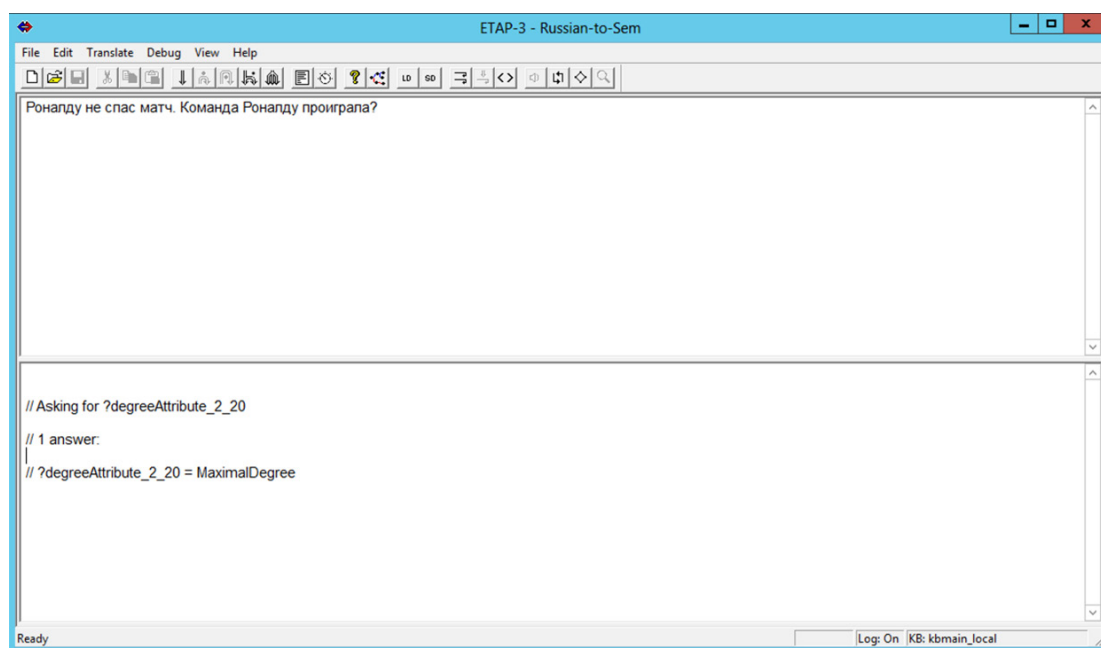


Fig. 2: Sentence (1) and the question *Команда Роналду проиграла?* 'Did Ronaldo's team lose the match?'

In the upper window of Figure 2 is the text (*Ronaldo did not save the match*) and the diagnostic question (*Did Ronaldo's team lose the match?*). The lower window contains the answer returned by the system. Let us explain this answer. For SemETAP, finding if a proposition is true amounts to discovering the value of the epistemic modality of this proposition. The meaning of the diagnostic question is: What is the value of the attribute `?degreeAttribute_2_20`, which is the value of the epistemic modality of the statement "the team for which Ronaldo was playing lost the match". In plain words, is it true that Ronaldo's team lost the match? The answer, which can be seen in the lower window, reads that the value of this modality is maximal. This means that the question is answered in the affirmative.

2. Ordinal Adjectives and Their Argument Structure

As mentioned above, semantic structures (SemS) constructed by the analyzer aim to represent the meaning of the sentence in an explicit and transparent way, while abstracting away from the lexico-syntactic variation of the NL. One important requirement for the SemS is that it should explicitly represent all the semantic arguments of the predicates of the sentence, expressed by all kinds of words – verbs, nouns, adjectives or adverbs (for more on the typology of different types of arguments see Boguslavsky, 2014). SemS is not constructed from NL words, but from semantic elements (concepts), which are connected by special semantic relations. Accordingly, the analyzer has two lexical resources – a semantic dictionary (ontology) and a combinatorial dictionary of Russian. All lexical information used by the analyzer to construct the SemSs is distributed between these two resources. The ontology contains an inventory of concepts and describes their properties – in particular, their argument structure. The combinatorial dictionary describes the properties of words and the relationship between these words and the ontology.

The lexical resources of the semantic analyzer should make it possible to solve the following tasks:

- a) Introduce the concepts (or combinations of concepts) necessary and sufficient for representing NL words and define their argument structure,
- b) Establish the correspondence between Russian words and ontological concepts expressing them in different contexts,
- c) Specify the ways in which each argument of these concepts is realized in the texts.

The first task is solved in the ontology, while the second and the third ones, in the combinatorial dictionary of Russian.

The subject of this paper are Russian ordinal adjectives and some related words (OAdj), which denote the position of an object in an ordered sequence of homogeneous objects, such as *first, second, third, ... last, penultimate, next, previous*. Their meaning is transparent enough. For example, adjective *пятый* 'fifth' in *пятая книга* 'the fifth book' means that the book under discussion occupies the fifth position in some ordered sequence of books. Surprisingly, the argument structure of these adjectives has not attracted due attention of either theoretical linguists or lexicographers of different languages and has not been described in the available literature and lexical resources with the necessary completeness (Geist, 2010; Grashchenkov, 2008; Grashchenkov & Kobozeva, 2017; Gutiérrez et al., 2005; Ikeya, 1995; Johnson & Fillmore, 2000; Meltzer-Asscher, 2011; Proceedings, 2022). In this paper we will try to fill this gap. It is obvious that in order to fully specify the position of an object in an ordered sequence, in a general case it is necessary to take into account several different aspects of the situation. For example, in the phrase *третий дом от аптеки* 'the third house after the pharmacy', the phrase 'after the pharmacy' signals where the sequence of numbered elements begins. The first element of this sequence is the house immediately following the pharmacy. Thus, several arguments (semantic slots) should be postulated for the concept of numbering objects and for the corresponding OAdj.

Below, we will enumerate the concepts proposed for OAdj and describe their argument structure. The ways in which arguments are expressed will be evident from the examples given. After that, we will comment on some non-obvious properties of OAdj.

The concepts used to represent OAdj are HavingOrderedPosition, Last, Penultimate, Next and Previous. All of them make part of the ontological class OrderingRelation. Below, we list their semantic arguments, indicating for each argument the semantic relation by which it attaches to the corresponding concept, a brief explanation, and examples.

Concept HavingOrderedPosition.

Arguments:

- **hasObject** – An entity having a number: *третий том* ‘third volume’, *пришел в третий раз* ‘came for the third time’.
- **belongsTo** – A sequence (or a domain) within which the object is thought of: *первая скамейка на бульваре* ‘first bench on the boulevard’, *седьмой в очереди* ‘seventh in the queue’, *третий альбом на верхней полке* ‘third album on the top shelf’, *второй дом по левой стороне улицы* ‘second house on the left side of the street’, *первый успех за 10 лет* ‘first success in 10 years’, *первая победа в турнире / из 10 матчей* ‘first win in a tournament/out of ten matches’; *второй раз за этот день* ‘second time this day’.
- **hasNumber** – The number of the object in this sequence.
- **orderedBy** – Criterion for ordering objects in a sequence: *третий по росту* ‘third in height’. This criterion may not be explicit. By default, events are ordered by time: *вторая победа* ‘second victory’ took place earlier than *третья победа* ‘third victory’. Physical objects are often ordered by their location relative to the observer: *третий дом по левой стороне* ‘the third house on the left hand side’ is closer to the observer than *четвертый дом* ‘the fourth house’.
- **hasStartingPoint** – Position from which starts the counting: *третий дом от аптеки (от конца)* ‘third house from the pharmacy (from the end)’, *третья строка снизу* ‘third line from the bottom’, *первый праздник после Нового года* ‘first holiday after the New Year’).
- **hasTerminalPoint** – Direction of counting: *пятая остановка в сторону центра* ‘fifth stop towards downtown’.

Here are some examples of SemS, which show how OAdj express their arguments in the text and how they are presented in semantic structures.

(2) *Третий дом от аптеки в сторону центра* ‘the third house from the pharmacy towards downtown’

House

```
isObjectOf HavingOrderedPosition
    hasNumber 3
    hasStartingPoint DrugStore
    hasTerminalPoint Downtown
```

For illustration purposes, this structure is visualized below by means of a graph:

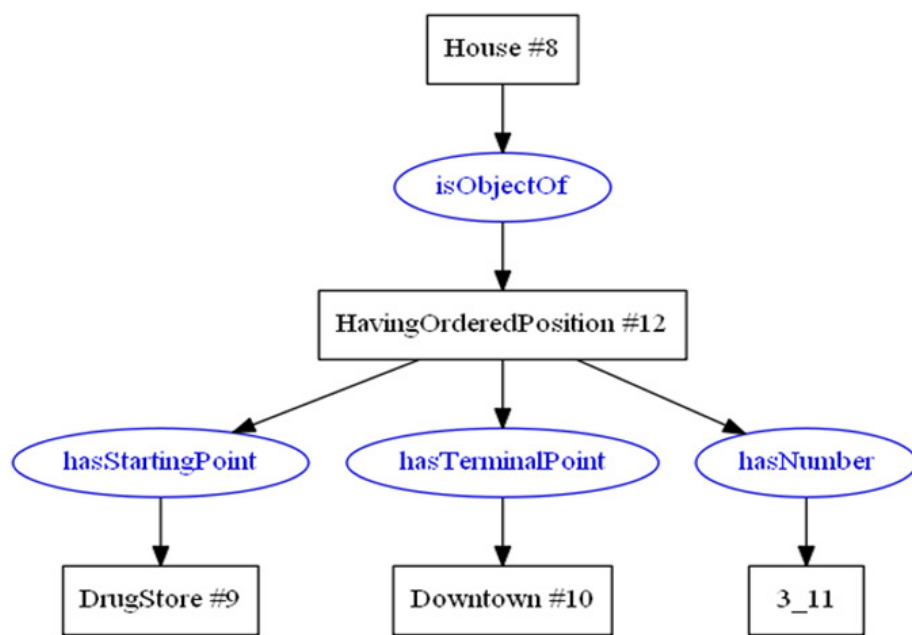


Fig. 3: The Basic SemS of the phrase *Третий дом от аптеки в сторону центра* ‘the third house from the pharmacy towards downtown’

(3) *Домик стоял четвертым слева на последней от штаба улице* ‘The cottage stood fourth on the left on the last street from the headquarters’.

HavingOrderedPosition

 hasObject House

 isObjectOf Location

 hasObject2 Street

 isObjectOf Last

 hasStartingPoint Headquarters

 hasNumber 4

 hasStartingPoint Left

(4) *Первый ученик в классе по росту* ‘the first student in the class in height (=the tallest)’

Student

 isObjectOf HavingOrderedPosition

 hasNumber 1

 belongsTo Class

 orderedBy HavingHeight

The Last and Penultimate concepts have the same arguments as HavingOrderedPosition except that they do not have the argument hasNumber. On the other hand, Next has a specific argument position (‘next after’ - hasObject2) absent in HavingOrderedPosition:

(5) *Экзамен, следующий в этой сессии по важности после математики* ‘the next most important exam in this session, after maths’.

Exam

isObjectOf Next

hasObject2 Mathematics

orderedBy Importance

belongsTo Session

3. Some Linguistic Peculiarities of OAdj Arguments

3.1 Starting Point of Numbering

The starting point argument shows where one should start counting from, but it may not be as simple as it seems. Let us compare sentences (6) and (7), which seem very similar:

(6) *Моя школа – второй дом после аптеки.*

‘my school is the second house after the pharmacy’

(7) *Вице-президент – второе лицо после президента.*

‘the Vice President is the second official person after the President’

In both cases, an object is specified which serves as a reference point for numbering (*the pharmacy, the President*), but there is an important difference. In (6), there are three houses before our eyes: the pharmacy, my school, and the house between the two. By contrast, in (7) there is no third person between the President and the Vice President. The former is immediately followed by the latter.

The difference in the interpretation of (6) and (7) is that in (6) the pharmacy is not thought of as a member of the numbered sequence. It only marks the point **after which the sequence begins**. Therefore, the house next to the pharmacy is numbered 1 in the sequence, and my school gets number 2. In (7), the President is **the element that starts the sequence**. Therefore, he gets number 1 and the next element (the Vice President) gets number 2.

Let us formulate the difference between (6) and (7) in more general terms. Let element A be *n*-th after X (in our examples (6)-(7) *n*=2). Then the meaning of (6) can be represented as follows. Element A (*my school*) has number *n* (2) in the sequence that begins after X (*the pharmacy*). Between X and A there are *n*-1 (1) unnamed elements of the sequence.

The meaning of (7) looks different. Element A (*Vice President*) has also number *n* (2) in a sequence, but X (*President*) is what constitutes the first *n*-1 (1) elements of the sequence.

The next difference between the above two variants of the reference point in (6) and (7) concerns the structure of the X group itself, that fills the hasStartingPoint slot. First of all, let us note that interpretation (7) requires a complete enumeration of all the first $n-1$ elements of the sequence. (8) is an example where $n=3$, and therefore X has to enumerate the first 2 elements of the sequence:

(8) *Спикер Палаты представителей Конгресса – третье лицо после президента и вице-президента.*

‘the Speaker of the House of Representatives of the Congress is the third person after the President and the Vice President’

In the next example $n=4$:

(9) *Россия является четвертым после Швеции, США и Дании государством по объему накопленных инвестиций* (Diplomatičeskij Vestnik, 2004).

‘Russia is the fourth country after Sweden, the USA and Denmark in terms of accumulated investment’

Let us also note that the elements of the coordinated phrase X cannot be switched. Sentence (9) not only tells us that Russia ranks 4th in terms of accumulated investment. It also tells us that Sweden ranks first, the United States second, and Denmark third. The coordinated elements should follow exactly the order in which the countries are ordered in terms of investment volume. Changing the order of the coordinated elements does not make the sentence grammatically incorrect, but it does make it false.

As far as sentences of the type (6) are concerned, they do not accept coordinated strings as hasStartingPoint arguments. Sentence (10) is ungrammatical (unless the pharmacy and the department store are situated in the same building).

(10) **Моя школа – второй дом после аптеки и универмага.*

‘my school is the second house after the pharmacy and the department store’

3.2 Numbered Sequence Includes the Main Predicate

As can be seen from the examples given, in a sentence OAdj arguments are usually expressed by groups syntactically subordinate to OAdj (with the exception of the numbered object itself, which is expressed by the noun being defined): *первый в классе* ‘first in class’, *второй по росту* ‘second in height’, *пятая строка снизу* ‘fifth line from the bottom’. Note, however, a case where the argument denoting the sequence itself (introduced by the belongsTo relation) includes the main predicate of the clause.

As a rule, OAdj together with their arguments are interpreted within the NP to which they belong. For example, in the sentence *Я купил билет в пятый вагон* ‘I bought a ticket to the fifth carriage’, the number of the carriage is defined without any connection with the verb *bought*. Meanwhile, there are sentences in which it is impossible to establish the sequence in which a given object is thought of without resorting to the main predicate of the clause. For example, the sentence (11)

(11) *Алькарас выигрывает уже третий турнир*

‘Alcaras wins (his) third tournament’

is natural to understand in the sense that Alcaras has already won two tournaments, rather than in the sense that in a series of somehow ordered tournaments he wins tournament number 3. In other words, the sequence in which the tournament in question has number 3 consists of the tournaments won by Alcaras.

4. Conclusion

The task of the semantic analyzer is to represent the meaning of a text in an explicit and transparent way. Identifying the arguments of all predicates of a text is the main subtask of this complex task, since predicate-argument relations are the “semantic glue” that unites individual semantic elements into a coherent semantic structure of a sentence. We consider the predicate-argument structure of a sentence somewhat more broadly than is often done. To the class of argument-containing words we include many adjectives and adverbs that have not been sufficiently studied in lexicography from this point of view. A good example is the compact class of ordinal adjectives (*first, second, third, ... last, next*). We show that in order to characterize the place of an object in an ordered sequence of similar objects it is necessary to take into account several parameters of the situation, such as the sequence within which the object is thought of (*the first bench on the boulevard*), the criterion for ordering objects in the sequence (*third in height*), the position from which the counting starts (*third line from the bottom*), and the direction of the counting (*fifth stop towards the city center*). As a rule, these parameters are expressed in the text by means of phrases syntactically connected with ordinal adjectives and meet all the requirements imposed on semantic valences (Mel’čuk, 2015). Some of these valences have interesting properties from the linguistic point of view (see Section 3).

References

- Boguslavsky, I., Frolova, T., Iomdin, L., Lazursky, A., Rygaev, I., & Timoshenko, S. (2018). Semantic analysis with inference: high spots of the football match. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue”*, 17(24), 124–142.
- Boguslavsky, I. M., Frolova T. I., Iomdin L. L., Lazursky A. V., Rygaev, I. P., & Timoshenko, S. P. (2019). Knowledge-based approach to Winograd Schema Challenge. In *Computational linguistics and intellectual technologies. Papers from the Annual International Conference “Dialogue”*, 18(25), 86–103.

- Boguslavsky, I. M., Dikonov, V. G., Frolova, T. I., Iomdin, L. L., Lazursky, A. V., Rygaev, I. P., & Timoshenko, S. P. (2020). Full-fledged semantic analysis as a tool for resolving Triangle-COPA social scenarios. In *Computational linguistics and intellectual technologies. Papers from the Annual International Conference "Dialogue"*, 19(26), 106–118.
- Boguslavsky, I. M., Dikonov, V. G., Frolova, T. I., Iomdin, L. L., Lazursky, A. V., Rygaev, I. P., & Timoshenko, S. P. (2023). Constructing a Semantic Corpus for Russian: SemOntoCor. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2023"*, 21(28), 1–14.
- Boguslavsky, I. (2014). Argument structure of adverbial derivatives in Russian. In J. Tsujii, & J. Hajic, *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers* (pp. 1071–1080). Dublin, Ireland, August 23–29, 2014.
- Geist, L. (2010). The argument structure of predicate adjectives in Russian. *Russian Linguistics*, 34, 239–260. <https://doi.org/10.1007/s11185-010-9064-5>
- Grashchenkov, P. V. (2008). *The Argument Structure of Russian Adjectives (ms.)*. Institute of Oriental Studies, Moscow, 37 p.
- Grashchenkov, P. V., & Kobozeva I. M. (2017). Semantic classes and government of adjectives. In Selegei et al. (Eds.), *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue"*, 16(23), 134–149.
- Gutiérrez, E., Javier, L., & Fábregas Alfaro, A. (2005). Cuestiones de orden: semántica y sintaxis de los adjetivos ordinales. *ELUA. Estudios de Lingüística*, 19, 85–105.
- Ikeya, A. (1995). Predicate-Argument Structure of English Adjectives. In B. K. T'sou, & T. B. Y. Lai, *Proceedings of the 10th Pacific Asia Conference on Language, Information and Computation*, (pp. 149–156), City University of Hong Kong, Hong Kong. City University of Hong Kong.
- Johnson C., & Fillmore Ch. (2000). The FrameNet tagset for frame-semantic and syntactic coding of predicate-argument structure. In *NAACL 2000: Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, April 2000, 56–62.
- Mel'čuk, I. (2015). *Semantics. From Meaning to Text. Vol. 3*. John Benjamins Publishing Company, Amsterdam/Philadelphia.
- Meltzer-Asscher A. (2011). *Adjectives and Argument Structure*. [Doctoral dissertation, Tel Aviv University].
- Proceedings of the Workshop on Aspect and Argument Structure of Adjectives/Adverbs and Participles/Prepositions*. (2022). University of Greenwich. https://docs.gre.ac.uk/__data/assets/pdf_file/0017/290042/waasap-10th-anniversary-conference-final.pdf

Acknowledgements

This work was supported by the RSF grant No. 24-18-00988, which is gratefully acknowledged.

Contact information

Igor Boguslavsky

Institute for Information Transmission Problems / Russian Academy of Sciences; Universidad Politécnica de Madrid.

Igor.M.Boguslavsky@gmail.com.

Nataliia Cheilytko and Ruprecht von Waldenfels

WORD EMBEDDINGS FOR DETECTING LEXICAL SEMANTIC CHANGE IN UKRAINIAN

Abstract The paper presents a count-based semantic vector space model for Ukrainian, which has been applied for the semantic change detection task. The approach assumes creation of multidimensional vector representations of occurrences for a particular lexeme or a group of related lexemes with further visual and quantitative analysis of the obtained semantic vector space. The multidimensional space has been reduced to 2D for visual data analysis with the Multidimensional Scaling technique. The paper described two case studies to show how the proposed R & D workflow helps revealing potential semantic change events and discuss benefits and limitations of the approach. One case study traces the disappearance of a regional sense, and another identifies the appearance of a new metaphoric sense that is widespread in the Ukrainian media discourse.

Keywords Ukrainian; corpus linguistics; GRAC; word embeddings; vector representation; semantic vector space model; semantic change; semasiological variation; multidimensional scaling; pointwise mutual information; semantic distance

1. Introduction

1.1 Goals and Motivation Semantic Change and Variation in Ukrainian

Ukrainian is a large European language with a high degree of variability and versatile semantic developments. This makes Ukrainian a perfect experimental field for exploring semantic trends in a systematic way, which is of high interest to both diachronic semantics and variationist sociolinguistics. Therefore, the paper focuses on a corpora-based empirical examination of semasiological changes in the Ukrainian lexicon within the XX–XXI centuries.

The project's research questions are: How and with what degree of certainty can we identify semantic change in Ukrainian lexemes given significantly large input textual data? What are the types of such changes? Are there any trends and hidden patterns in how semantic changes are distributed across different regions of the country within various registers and domains?

The current paper looks into two specific extralinguistic circumstances in Ukraine's history that triggered significant instances of semantic change and lexical variation to assess the proposed methodology. The first case is the fusion of two separate Ukrainian standards after WWII due to the unification of the Ukrainian territories within the Soviet Union. Extralinguistic cataclysmic events often significantly influence language development, as language users revisit conceptual representations

under such circumstances. The second case is therefore the Russian full-scale invasion of Ukraine that started on February 24th, 2022.

1.2 Corpus Data

Data-wise, the research uses the concordances from the following corpora: the General Regionally Annotated Corpus of Ukrainian (GRAC), which represents Ukrainian starting XIX century and up to date in various domains and regions, including diaspora (Shvedova et al., 2017–2024), and the Ukrainian Trends corpus available within the Sketch Engine platform for the most recent data in the Ukrainian digital media. For the first case study, we used textual data from the three periods: 1920–1939 (approx. 10M tokens), 1940–1969 (approx. 30M tokens), 1970–1990 (approx. 50M tokens). The data included from fiction and periodicals and represented texts published in both the Western and the Eastern region of Ukraine. For the second case study, we considered two sub-corpora with texts from 1960–2019 and 2020–2024 from fiction, periodicals, and media texts. All texts have been processed with the help of the Ukrainian Part-of-Speech tagger and lemmatizer developed by Andrii Rysin¹.

1.3 Semantic Vector Space Representation

Our research is methodologically grounded on the distributional hypothesis that posits that one can distinguish the meanings of a word by their typical occurrences and collocates (Schütze, 1998). For example, in the following two sentences: “*I have a great fan of rock among my friends*” (fan as a person) and “*I have a great fan with several heating options*” (fan as a cooling, heating device), we can see the difference in the senses relying on the contextual surrounding (*rock, friends* vs. *heating options*). In order to identify and explore semantic variation and change from a usage-based perspective and in a bottom-up fashion by performing both quantitative and qualitative comparison of word occurrences, we construct and analyze semantic vector space representations (aka semantic word embeddings) of tokens following the research pipeline proposed by (Hilpert & Correia Saavedra, 2020), (Montes & Geeraerts, 2022) and (Geeraerts et al., 2023). Vector space models treat text as a multidimensional vector space, where a word, a token as a particular utterance of a word, or a combination of words, or a sentence are represented as a vector so that it is possible to apply various vector algebra operations: calculate distance between vectors, apply dimensionality reduction to a vector space, visualize and cluster them.

The calculation of vector representations is often grounded on co-occurrence patterns of a particular linguistic entity (a word, a token, etc.). It considers how frequently other linguistic entities are being used in a particular corpus within a particular surrounding of the target word (Turney & Pantel, 2010, p. 149).

In our research, we build vector representations for tokens and account not only for the immediate surrounding collocates but also for second-order collocates – typical collocates of words found in concordance lines built for a given token of interest. In

¹ https://github.com/brown-uk/nlp_uk

such a way, a word occurrence vector includes information not only about collocates observed in its surroundings, but also collocates of the collocates. In other words, if to follow a social network metaphor, one can be distinguished not only by their friends but also by friends of a friend (Hilpert & Correia Saavedra, 2020, p. 394).

1.4 The Semantic Vector Space Pipeline

The text processing and analysis pipeline contains the following steps:

1. **First-order vectors.** For a given corpus, calculate a co-occurrence matrix with first-order type-based vectors for the vocabulary of the 10,000 most frequent words in the corpus, with the exception of the most frequent grammatical words (prepositions, conjunctions, and the likes). Normalize the matrix with the Pointwise Mutual Information (PMI) index, keep only those columns and rows in the matrix that contain at least one PMI no less than 3.
2. **Second-order vectors.** Build a second-order token-level vector representation for each utterance of a lexeme:
3. Get first-order vectors from step 1 for each collocate found in the surrounding of the lexeme under examination;
4. Average those vectors so that each lexeme utterance gets its averaged vector that accounts for second-order collocates found in the corresponding utterance of a particular size.
5. Apply **multidimensional scaling (MDS)** technique to the second-order embeddings representing utterances of a particular lexeme under examination and **visualize semantic vector space** for the occurrences as 2D scatterplots as proposed in (Wheeler, 2005), where each dot stands for a second-order vector of a single utterance. Perform visual analysis of the plots to see trends in lexeme usage – across time, regions, and registers.
6. Apply MDS for the lexeme in question with specific **anchor word(s)** to see how lexeme's second-order embeddings are located in a 2D semantic vector space compared to the embeddings built for the anchor word(s). For example, in section 3 of this paper, for the Ukrainian noun *bavovna* 'cotton', we use the anchor word *l'on* 'lien, flax'.
7. Calculate the **average Euclidean distance** within a group of token embeddings of a lexeme within a certain period, as well as average distances to a similar group of token embeddings but of a different period and/or register so that a significant change of the distance may indicate a change in how the target lexeme is being used.

This pipeline was sufficiently successful in identifying evidence for semantic changes here. However, the approach faces complications and uncertainties in certain cases, which are discussed below.

2. Semantic Change in the Context of Merging Standard Ukrainian

2.1 Ukrainian as a Multi-Standard Language Before WWII

We consider Ukrainian to be a multi-standard language with significant variation and fast language development due to intense and active social and political changes. As shown in (Shvedova, 2021) and (Lahjouji-Seppälä et al., 2022), in the time before WWII Ukrainian had two standard language varieties, as the Ukrainian-speaking territory was historically separated between two neighbouring countries – between Habsburg, then Poland in the West and the Russian Empire and the Soviet Union in the East. The two standards were merged after WWII when the Ukrainian territories were united within the Soviet Union. As a result, a similar language policy was applied to the whole country. Due to this and extensive ethnic cleaning, the influence of the Polish language contact became much less significant.

In order to perform a systematic bottom-up exploration of how those processes influenced semantic variation and change in Ukrainian, we applied the R & D pipeline briefly described in the previous section.

2.2 Disappearance of a Word Meaning

This case study presents how the adjective *povazhnyi*, meaning ‘respectable, honorable’ and also ‘serious, severe’, has almost lost the latter sense during the second part of the XX century due to the fusion of the standard Ukrainian varieties. Figure 1 contains the 2D visualization of the semantic vector space obtained with the multidimensional scaling for regionally annotated occurrences of this adjective (K – Kyiv representing the Eastern part of the country, L – Lviv representing the Western part). The plot in Figure 1 shows the complete overlap between the two regional groups of embeddings built for adjective occurrences in 1970–1990. Further qualitative analysis of the occurrences confirmed that the specific Western sense ‘serious, severe’ is almost absent for both regions, and thus, such overlap indicates that this sense is disappearing.

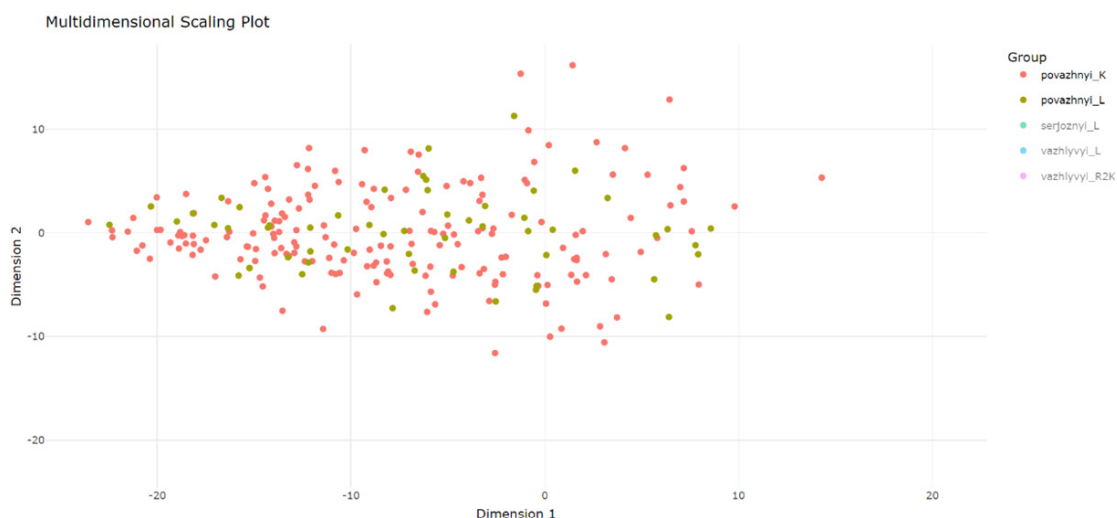


Fig. 1: Embeddings of the *povazhnyi* occurrences in 1970–1990 (pink – Eastern Ukraine, brown – Western)

Figure 2, on the contrary, demonstrates that before WWII (1940–1969), the Western occurrences have their own area on the plot (the central lower part of the plot) where the Eastern occurrences are not present. Indeed, manual inspection of the occurrences in question shows that this regionally specific area most often involves cases with the ‘serious, severe’ sense, which is much less usual for the other area on the plot. In this way, the visual representation of embeddings indicates regional differences in usage.

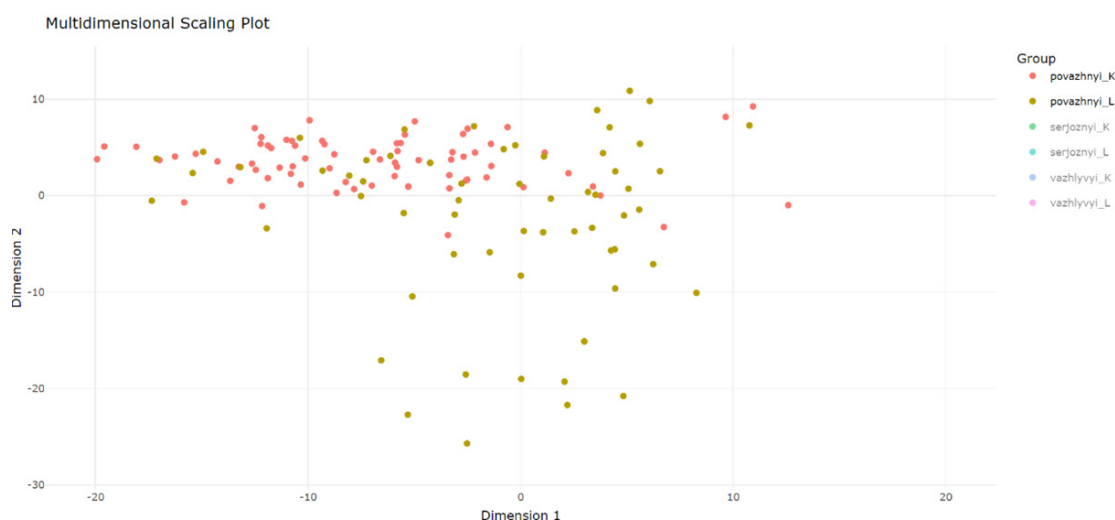


Fig. 2: Embeddings of the *povazhnyi* occurrences in 1940–1969 (pink – Eastern Ukraine, brown – Western)

The same observation is valid also for an earlier period of 1920–1939 (Figure 3): the area on the plot where the data from the two regions do not overlap reflects cases where the adjective is used in the sense ‘serious, severe’ – most of which are from the Western region.

We conclude that the adjective *povazhnyi* had a regional sense ‘serious, severe’ in West in the period when Ukrainian was a multi-standard language. We explain this with close and strong linguistic contact with Polish in West, where the cognate the adjective *powazhny* primarily denotes ‘serious, severe’. After the war, this sense almost disappeared from use due to the general trend of Ukrainian to become closer to its Eastern standard in this period.

It is important to mention that the proposed case study is just a single example from a systematic lectometric assessment we are performing for standard Ukrainian targeted to show how restructuring of a multi-standard Language results in semantic structure changes of a word or a concept.

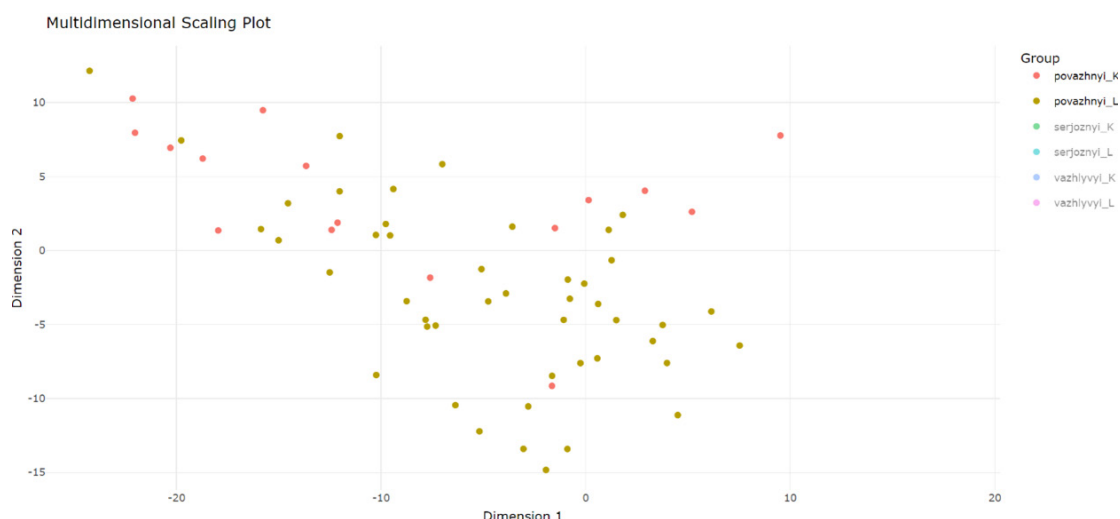


Fig. 3: Embeddings of the *povazhnyi* occurrences in 1920–1939 (pink – Eastern Ukraine, brown – Western)

3. Semantic Change in Cataclysmic Events

3.1 Appearance of a New Metaphoric Sense

With the MDS-based visualisation of second-order collocation vectors of a particular lexeme done for different periods, it becomes possible to identify a significant change in how words are used, and reveal new meanings. Figure 4 demonstrates the vector space for the Ukrainian noun *bavovna* ‘cotton flower, cotton material’ for two periods: 1960–2019 and 2020–2024 – before and during the ongoing war in Ukraine. The plot shows that most of the occurrences of the same word do not overlap but are located in different areas in the vector space, which is a strong indicator of semantic change. Indeed, this noun has acquired a completely new metaphoric sense ‘explosion’, since the word is being used as a euphemism referring to the results of a fire attack.

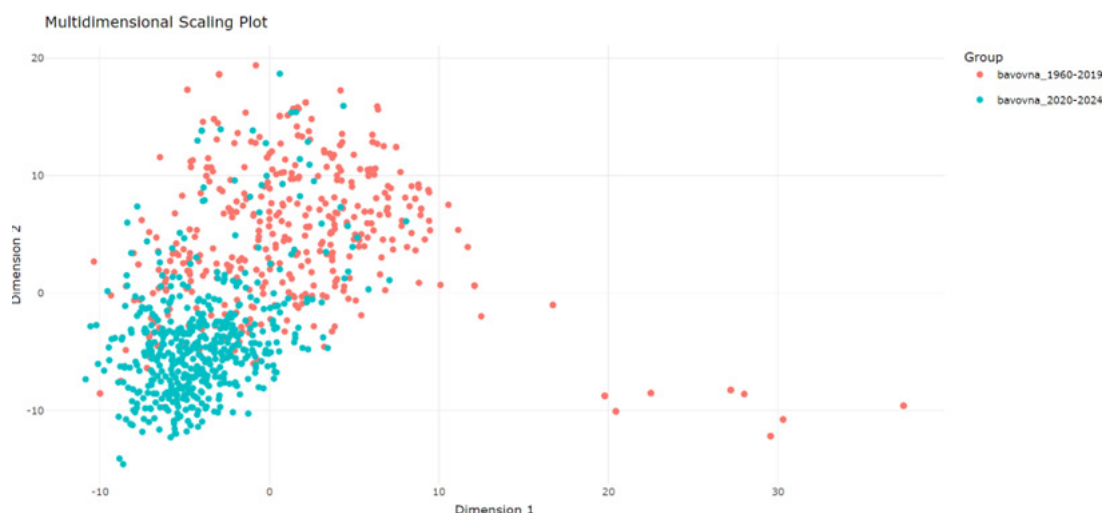


Fig. 4: Embeddings of the *bavovna* occurrences (pink – 1960–2019, blue – 2020–2024)

This usage originated from a wordplay with Russian *hlopok*, a homograph for two words meaning ‘cotton’ (same as Ukrainian *bavovna*) and ‘a clap’, implying an undisturbing and quiet sound, which Russian media often use to disguise the facts of fire attacks on Russian military and infrastructure objects, so that to avoid panic among their citizens. The Ukrainian media confronting and mocking this fact-hiding strategy of the opponents uses the Ukrainian equivalent of ‘cotton’, i.e., *bavovna*, to denote events of explosions caused by fire attacks. As a result, the noun has acquired a new metaphoric sense ‘explosion’ in Ukrainian media texts. In other registers, however, the noun continues to refer to the classical meanings ‘a cotton flower’ and ‘cotton material’.

3.2 Anchor Words to Explore Semantic Change

A further technique to identify and examine evidence of semantic change is to build a vector space plot for a target word with certain anchor words (near-synonyms, antonyms) to see how the target word is being used in comparison to these other words over time. Figure 5 explicitly shows that before the Russian invasion (1960–2019), the near-synonyms *bavovna* ‘cotton flower’, ‘cotton material’ and *l'on* with the meanings ‘flax flower’, ‘linen material’, and ‘flax seeds as a food product’ have significantly larger overlapping area, so they were used as near-synonyms. With the invasion, however, the dots on Figure 6 representing the *bavovna* embeddings become more concentrated on the left part of the semantic vector space and almost do not overlap with the *l'on* embeddings. This differentiation from a word with a similar meaning is another indication of the semantic change.

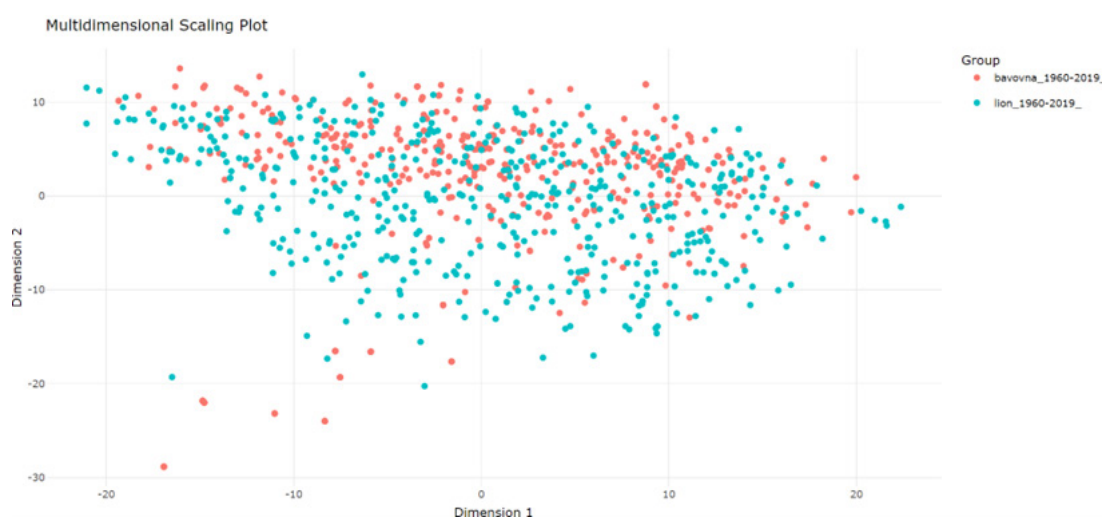


Fig. 5: Embeddings of the *bavovna* (pink) and *l'on* (blue) occurrences in 1960–2019



Fig. 6: Embeddings of the *bavovna* (pink) and *l'on* (blue) occurrences in 2022–2024

Similarly, Figures 7–8 demonstrate how *bavovna* relates to the noun *vybuh* ‘explosion’. Before the Russian invasion (Figure 7), the embeddings of those two words take mostly distinct areas on the semantic vector space. In contrast, after the invasion, their embeddings significantly overlap (Figure 8), proving that these words start sharing their usage patterns and thus meaning – ‘explosion’.

Therefore, the distributional methodology grounded on the second-order vector representations helped to explore how particular lexemes function over time and across different regions and registers and, thus, to reveal potential indications of conceptual reorganization in the Ukrainian lexicon.



Fig. 7: Embeddings of the *bavovna* (pink) and *vybuh* (blue) occurrences in 1960–2019

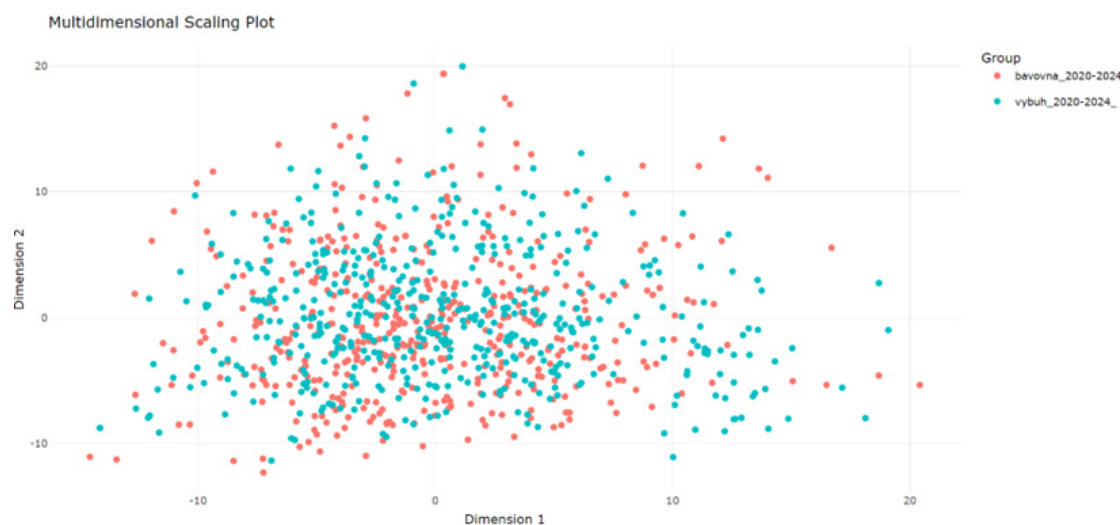


Fig. 8: Embeddings of the *bavovna* (pink) and *vymbuh* (blue) occurrences in 2020–2024

4. Moving on: LLM Representations vs. Count-Based Vectors

The current approach is based on the semantic vector space model obtained by counting frequencies of collocates in a contextual window for a target word. Such models are considered to be good from the explanatory perspective. However, they have several limitations, among which is the inability of a single generic count-based vector space model to represent the entire lexicon as used in a multitude of registers, and therefore, a plurality of models would be required for a systematic exploration of language change. Moreover, such models face certain complications with the words-sense disambiguation task, which is an essential step of any semasiological modelling, as collocations are indicative on many other factors besides word senses.

To overcome the limitations in the next phase of our research, we are going to expand the solution by adding vector representations (embeddings) obtained from different transformer models fine-tuned for Ukrainian (Laba et al., 2023). Our preliminary test study showed that ChatGPT-4o is able to distinguish various senses for *bavovna*, including the new metaphoric sense, as well as for a group of other Ukrainian polysemous words.

5. Conclusions

The second-order vector representations, together with the MDS-based visualisation technique, make it possible to identify events of semantic change in Ukrainian across different periods, registers and geographic regions. The paper presented two case studies to demonstrate the research workflow for exploring such changes. The approach is being used for a large-scale bottom-up systematic exploration of lexical variation and change in Ukrainian.

In order to further tune and assess the methodology, the authors are preparing a representative test dataset with known cases of Ukrainian variation and semantic

change. However, the preliminary goal is to monitor and identify unknown cases and, therefore, to reveal hidden patterns and trends in Ukrainian language development.

A systematic study of linguistic variation and semantic change in Ukrainian will contribute to the fundamental understanding of universal mechanisms of language development by comparing the findings to similar studies done for other multi-standard languages like German, Dutch, English, French, Spanish, and Portuguese.

Apart from that, such a consistent dynamic large-scale view on semantic evolution may bring additional insights with respect to neighbour language contact influence, in the case of Ukrainian, with Polish and Russian.

Last but not least, the proposed R & D pipeline is not language-specific and can be applied to various languages.

References

- Geeraerts, D., Speelman, D., Heylen, R., Montes, M., de Pascale, S., Franco, K., & Lang, M. (2023). *Lexical Variation and Change. A Distributional Semantic Approach*. Oxford Academic.
- Hilpert, M., & Correia Saavedra, D. (2020). Using token-based semantic vector spaces for corpus-linguistic analyses: From practical applications to tests of theoretical claims. *Corpus Linguistics and Linguistic Theory*, 16(2), 393–424. <https://doi.org/10.1515/cllt-2017-0009>
- Laba, Y., Mudryi, V., Chaplynskyi, D., Romanyshyn, M., & Dobosevych, O. (2023). Contextual Embeddings for Ukrainian: A Large Language Model Approach to Word Sense Disambiguation. In M. Romanyshyn (Eds.), *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)* (pp. 11–19). Association for Computational Linguistics.
- Lahjouji-Seppälä, M. Z., Rabus, A., & von Waldenfels, R. (2022). Ukrainian standard variants in the 20th century: stylometry to the rescue. *Russian Linguistics*, 46(3), 217–232. <https://doi.org/10.1007/s11185-022-09262-9>
- Montes, M., & Geeraerts, D. (2022). How vector space models disambiguate adjectives: A perilous but valid enterprise. *Yearbook of the German Cognitive Linguistics Association*, 10(1), 7–32. <https://doi.org/10.1515/gcla-2022-0002>
- Schütze, H. (1998). Automatic word sense discrimination. *Computational Linguistics*, 24(1), 97–124.
- Shvedova, M., von Waldenfels, R., Yarygin, S., Rysin, A., Starko, V., & Nikolajenko, T. (2017–2024). *GRAC: General Regionally Annotated Corpus of Ukrainian*. Retrieved July 31, 2024, from <https://uacorporus.org/>
- Shvedova, M. (2021). Lexical Variation in the Language of the Ukrainian Press of the 1920s–1940s and the Development of a New Lexical Norm. *Movoznavstvo*, 1, 16–35. <https://doi.org/10.33190/0027-2833-316-2021-1-002>
- Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37, 141–188. <https://doi.org/10.1613/jair.2934>

Wheeler, S. (2005). Multidimensional scaling for linguistics. In R. Köhler, G. Altmann, & R. G. Piotrowski (Eds.), *Quantitative linguistics. An international handbook* (pp. 548–553). Walter de Gruyter.

Contact information

Nataliia Cheilytko

Friedrich Schiller University Jena
natalia.cheilytko@gmail.com

Ruprecht von Waldenfels

Friedrich Schiller University Jena
ruprecht.waldenfels@uni-jena.de

Lian Chen, Wenjun Sun, and Flora Badin

INNOVATION IN PHRASEOMATICS

DiCoP Project and DiCoP-Text Corpus for the Enrichment of Language Models and Automatic Translation

Abstract This article examines advances in phraseomatics (Chen, 2023) and digital phraseography through the DiCoP project and its DiCoP-Text corpus, aimed at enriching linguistic models and machine translation. The project evaluates the frequency of use of phraseological units (PUs) and improves their translation in different contexts, drawing on recent research in phraseotranslation (Sułkowska, 2022) and natural language processing (NLP). It emphasizes French-Chinese and Chinese-French language pairs. We integrated 549 PUs from the novel *The Three-Body Problem* by Liu Cixin for our tests. Various processes, such as tokenization, identification, alignment, and annotation, were used to improve the translation of PUs. DiCoP-Text, a comprehensive database including newspaper articles, literary works, and textbooks, aims to enhance the performance of language models (LMs).

Keywords DiCoP; DiCoP-Text; phraseomatics; phraseotranslation; digital phraseography; NLP; language models

1. Introduction: Digital (Meta)Phraseography and Phraseomatics

In the contemporary digital age, the field of computerized phraseography faces challenges, which highlights the emergence of a particular discipline: phraseomatics or computational phraseology (Chen, 2023). This article explores the intricacies between computer science and the analysis of phraseological units or PUs (Gonzalez-Rey, 2002; Bolly, 2011, p. 28) to understand the subtleties of semantics and the structure of expressions in a digital context. This emerging discipline highlights the challenges and opportunities inherent in the evolution of (meta)phraseography (Chen, 2022; 2023).

In this sense, we present the Dictionary and Corpus of Phraseology (DiCoP)¹ project we are currently developing. The main objective is to create an electronic dictionary of multilingual phraseology (currently French-Chinese and Chinese-French) concerning PUs. Thus, this article is concerned with Chinese into French and, more specifically, among the innovative aspects of this project, the DiCoP-Text, which will serve as a corpus for phraseotraduction.

Indeed, in French and Chinese, PUs are ubiquitous in daily use (Bolly, 2011), so their recognition poses a major challenge in natural language processing (NLP). The DiCoP project addresses this problem by creating a bilingual electronic dictionary

¹ To learn more about the architecture of the macrostructure and microstructure of this innovative DiCoP digital dictionary, we invite you to consult our work presented at the Asialex conference (Chen, 2023).

(with multilingual expansion envisaged) and developing a corpus of common PUs (e.g., collocations, idiomatic expressions, and proverbs). However, identifying PUs in a monolingual source text and translating them correctly for a parallel (bilingual) corpus constitutes a real challenge for automatic phraseotranslation (Sułkowska, 2022; Chen, 2022), especially since these expressions are numerous and possess metaphorical and opaque semantics (Gross, 1996).

2. Corpus of Phraseotraductology: DiCoP-Text

It is possible to determine corpus types depending on the characteristics of a corpus and the origin of its data. Bowker and Pearson (2002) distinguished the following corpus types: general/specialized, synchronic/diachronic, written/oral, and monolingual/bilingual. Monolingual corpora are comprised of data from just one language, while multilingual corpora are comprised of data from several languages and are either parallel or comparable.

DiCoP-Text refers to a collection of texts used to study PUs. This corpus contains a comparative component, in each language separately, sourced from various media (e.g., literary works, poetry, magazines, and newspapers) to analyze the usage and check the vitality (e.g., type and frequency) of PUs in a specific language. It also includes a parallel component (i.e., bilingual, primarily sourced from literary works, as other media lack existing translations) to enhance bilingual translations.

The collection and digital processing of texts are currently undergoing expansion. To illustrate the application of NLP in the DiCoP-Text project and analyze PUs, more precisely, *chéngyǔ*² corresponding to French idiomatic expressions, we selected a parallel corpus: the novel *The Three-Body Problem* (三体 *Sāntǐ*) by Liu Cixin, with its French translation. The novel comprises 186,079 words in the original Chinese version and 141,279 in the French version.³

3. DiCoP-Text and NLP: Analysis of Parallel Corpus Results

To develop the corpus and NLP, we started by processing various electronic or digitized file formats, converting them to conventional formats such as .txt. The corpus is designed to ensure its quality and relevance. We then analyzed the data to extract morphosyntactic and syntactic information, and automatically selected PUs from the corpora. We will now present our method and the steps of processing the corpus in a systematic and rigorous manner.

² “*Chéngyǔ* are polylexical sequences, fixed phrases, or short sentences that function as monolexical units within a sentence. Semantically, they are endowed with a specific, non-compositional meaning that cannot be directly deduced from the individual characters. Syntactically, their basic form, which most often follows a fixed quaternary (quadrisyllabic) rhythm and is divided phonetically and/or syntactically into two hemistichs, is conventional and unchanged for generations; hence the name *chéngyǔ*, meaning “ready-made expressions.” Culturally, they convey the idiosyncrasies of a culture. Most often derived from classical literary language, they reflect an elegant and concise style and frequently contain strong allusive content” (Chen, 2021, p. 129). For example: 佛口蛇心 *fókǒu-shéxīn* (Buddha’s mouth, serpent’s heart): “the mouth of a Buddha, the heart of a snake” (Prov.).

³ For example, in French, “avoir la tête dans les nuages” (have your head in the clouds).

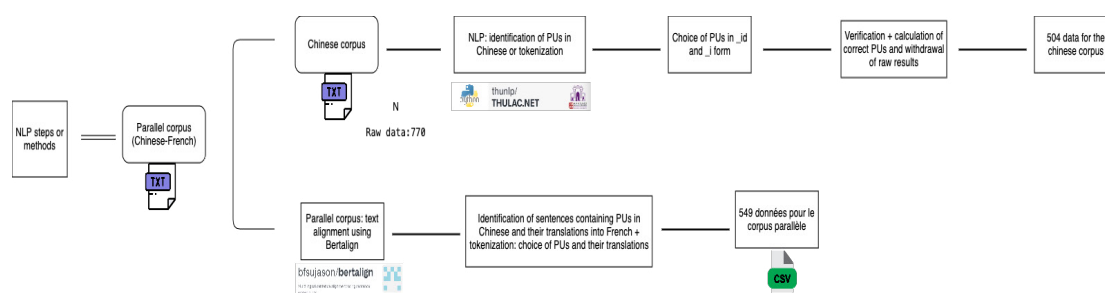


Fig. 1: NLP process followed for the parallel corpus

3.1 Performance and Evaluation of PU Identification in the Chinese Version with Thulac

For the Chinese analysis, we used the Thulac⁴ tool with Python to identify *chéngyǔ*, which we marked as `_i` or `_id` to indicate they were idioms. Our results are described below.

Table 1: Chinese monolingual corpus performance and evaluation results

	Chinese PUs identified by Thulac			
Total	771			
Success rates	65.43% (504/771)			
Error rate	34.57% (267/771)			
	Form <code>_id</code> (567 PUs)	Form <code>_i</code> (204 PUs)		
	Correct	Error	Correct	Error
Number of IEs	321	246	183	21
Percentage	56.61 %	43.39%	89.71%	10.29%

We identified 771 instances marked as `_id` or `_i`. Of these, 504 were confirmed correct, with 267 errors not constituting PUs. Even with a success rate of approximately 65.43%, this tool greatly facilitated identification and tokenization at this stage. We also checked 66 idiomatic expressions (IEs) that the Thulac tool did not identify as PUs.

Furthermore, we observed repetitions in the IEs listed in this table. For example, the expression 一动不动 *yīdòngbùdòng* appeared seven times but was translated differently depending on the case, such as “ne bougeait plus” (no longer moved), “être immobiles” (to be still), and “resta immobile” (remained still), or it could be omitted in the translation depending on the specific context.

3.2 Data Alignment in Chinese and French

Data alignment in Chinese and French was conducted using the Bertalign⁵ tool, which is known for its effectiveness in multilingual alignment. Bertalign

⁴ <http://thulac.thunlp.org/>

⁵ <https://github.com/bfsujason/bertalign>

achieves more accurate results than the traditional length-, dictionary-, or MT-based alignment methods (Thompson & Koehn, 2019). Bertalign identifies correspondences between words in both languages and facilitates the subsequent extraction of Chinese PUs and their French translations. Hence, the tool helps avoid literal translations of compositional expressions, which is a key challenge in machine translation.



Fig. 2: Each sentence in Chinese and its French translation after alignment

After alignment, the results were evaluated for accuracy by analyzing selected sentences to guarantee the quality and reliability of the aligned data. Sentences containing these idioms and their translations were compiled in a CSV file for detailed analysis. By focusing on sentences containing PUs, we provided essential enriched context to enhance the efficiency of automatic translation. This targeted approach, recommended by Artetxe et al. (2018), eliminated the need for large parallel corpora. The PUs identified were associated with equivalents in the target languages and were classified by a confidence value.

These data were then passed on to our engineers for technical integration and optimization, ensuring efficient linguistic data management in a multilingual context.

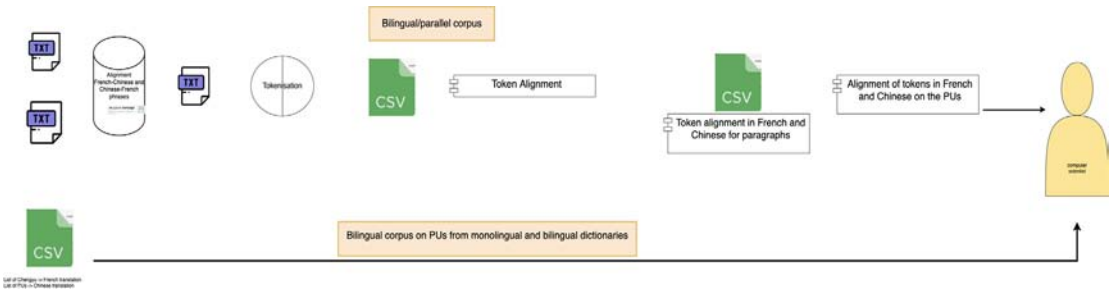


Fig. 3: Processes from NLP to IT

4. DiCoP-Text for the Enrichment of Language Models and Automatic Translation: Phraséo – Tratuctologie et Informatique

Today, neural network technology is widely used in machine translation task, especially pre-trained models methods. Pre-trained models are trained with large-scale datasets and then fine-tuned for specific subtasks to achieve better performance. However, because of the dependency on pre-trained models for data and the lack of specific PU training data, the current pre-trained translation models faced challenges when the sentences contained PUs.

Thus, we proposed an approach to improve the translation of machine translation models' PUs.⁶ First, we updated the vocabulary of the tokenizer of the pre-trained model based on PUs and then renewed the model's embedding layer so that the model could recognize PUs from the input context and produce the new token embedding. Then, we fine-tuned the model with individual PU data to improve its ability to embed the PUs into a semantic vector.

We proceeded to sentence-level training after the model could distinguish and embed individual PUs. We split the sentence data into training, validation, and test sets. With this idea, we tested if fine-tuning with the PU corpus could improve the machine translation model's ability to recognize and translate PUs at the sentence level.

Experiment, Results, and Analysis

We first extracted Chinese PUs from all datasets and compared each PU with the tokenizer of the language model. We added it if there was no corresponding item in the tokenizer's vocabulary while expanding the language model's embedding layer to initialize the weights of the added Chinese PUs. In the training phase, the model learned the embeddings of all Chinese PUs and the training sentence set. Then the model was tested on the test sentence set.

We selected these language models:⁷ Mbart (Tang et al., 2020), M2m100 (Fan et al., 2021), Nllb (Costa-jussà et al., 2022), and Mrebel (Cabot et al., 2023). For metrics, we chose SacreBLEU (Post, 2018). The experiment corpus had 549 data items. We divided the corpus into training, validation, and test sets at a ratio of 6:2:2. Subsequently, 409 new tokens were added for each of the four language models to the tokenizer. In the training phase, we set learning to 4e-5 and the batch size to 8. For the translation task, we selected Chinese to French.

The results are shown in Table 2. The "Original model" and the "Fine-tuned model" designate the performance of the original and fine-tuned language models, respectively.

⁶ The relevant code is available at the anonymous code repository: <https://anonymous.4open.science/r/DiCo-C7BA>

⁷ The weights files of all the language models can be downloaded at this address separately: <https://huggingface.co/Babelscape/mrebel-large>, <https://huggingface.co/facebook/mbart-large-50-many-to-many-mmt>, <https://huggingface.co/facebook/nllb-200-distilled-600M>, and https://huggingface.co/facebook/m2m100_418M.

Table 2: The experimental results

Model	Original model	Fine-tuned model	Improved value
Mbart	2,0686	16,7191	14,6505
M2m100	4,9078	16,2120	11,3042
Nllb	4,6345	17,1385	12,5040
Mrebel	0,0131	7,1798	7,1667

Based on the experimental results, we noted that after fine-tuning, all language improved in translation performance, indicating that the introduction of PUs can help translation models better understand PUs. Although each fine-tuned model outperformed its original model, they all scored low. This phenomenon was related to the volume of the training data, as the pre-trained models needed a large corpus to enable them to learn as much as possible about the semantics of individual tokens under different contents. However, the dataset used in this experiment had only 549 items, which was insufficient. Therefore, the fine-tuning performance of the model will be further improved after the PUs corpus is expanded. Nonetheless, fine-tuning could not completely make up for the shortcomings of the original model. In addition to expanding the PU corpus, using the Chinese-French translation corpus of ordinary texts to improve the Chinese-French translation ability of the model can also improve the performance. We will use methods such as prompt engineering to explore the effects of corpus improvement in future work.

5. Conclusion

The first evaluation of the NLP tools in the DiCoP-Text project provided a detailed overview of the effectiveness of the DiCoP-Text corpus and the improved LMs. Our proposal aimed to improve LMs by integrating more fixed expressions and refining linguistic models for more accurate identification and translation of PUs. However, room for improvement exists:

- 1) In the future, we envision broader applicability of our DiCoP project. Indeed, expansion to other languages would strengthen its relevance and applicability across the IT language community. We also aim to provide more information, including details concerning user interfaces, accessibility, and integrating user feedback into ongoing development.
- 2) We studied the effect of fine-tuning the translation model using a PU corpus. This approach involved updating the tokenizer, training the model to integrate these PUs, refining the model from sentences containing them, and testing based on sentence-level data. The results indicated that fine-tuning the model with PUs could improve its translation capacity. However, given the limitations of the model and the corpus volume, additional efforts are necessary to refine its translation capacity. Thus, our future research will focus on expanding the corpus and improving the model's Chinese-French translation capability.

References

- Artetxe, M., Labaka, G., & Agirre, E. (2018). A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In I. Gurevych, & Y. Miyao (Eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 789–798). Association for Computational Linguistics.
- Bolly, C. (2011). *Phraséologie et collocations. Approche sur corpus en français L1 et L2*. Peter Lang.
- Bowker L., & Pearson, J. (2002). *Working with Specialized Language – A Practical Guide to Using Corpora*. Routledge.
- Cabot, P.-L. H. et al. (2003). REDFM: A Filtered and Multilingual Relation Extraction Dataset. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 4326–4343). Association for Computational Linguistics.
- Chen, L. (2023). Meta)phraseography and phraseomatics: DiCoP, a computerized resource of phraseological units. *Proceeding of ASIALEX 2023: Lexicography, Artificial Intelligence, and Dictionary Users – The 16th International Conference of the Asian Association for Lexicography* (pp. 224–231). Asian Association for Lexicography.
- Chen, L. (2022). Phraséoculturologie: une sous-discipline moderne indispensable de la phraséologie. *SHS Web of Conferences – 8e Congrès Mondial de Linguistique Française – CMLF 138* (04011), 1–18. <https://doi.org/10.1051/shsconf/202213804011>
- Chen, L. (2021). *Analyse comparative des expressions idiomatiques en chinois et en français (relatives au corps humain et aux animaux)*. Doctoral dissertation, University of Cergy Paris]. HAL.
- Costa-jussà M.-R. et al. (2022). *No language left behind: Scaling human-centered machine translation*. Retrieved July 30, 2024, from <https://arxiv.org/pdf/2207.04672>
- Fan, A. et al. (2021). Beyond English-centric multilingual machine translation. *Journal of Machine Learning Research*, 22 (107), 1–48.
- Gross, G. (1996). *Les expressions figées en français: noms composés et autres locutions*. Ophrys.
- González-Rey, M. I. (2002). *La phraséologie du français*. Presses Universitaires du Mirail.
- Jha, A., & Patil, H. Y. (2023). A review of machine transliteration, translation, evaluation metrics and datasets in Indian Languages. *Multimedia Tools and Applications*, 82(15), 23509–23540.
- Lu, Y., Zeng, J., Zhang, J., Wu, S., & Li, M. (2021). Attention calibration for transformer in neural machine translation. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 1288–1298). Association for Computational Linguistics.

Mejri, S. (2013). Figement et défigement: problématique théorique. In L. Perrin (Eds.), *Pratiques: Le figement en débat*, n° 159–160, 79–97.

Mel'čuk, I., & Polguère A. (2007). *Lexique actif du français*. De Boeck.

Nelson, M. (2010). Building a written corpus. In A. O'Keeffe, & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 53–65). Routledge.

Post, M. (2018). A Call for Clarity in Reporting BLEU Scores[C]. In O. Bojar et al. (Eds.), *Proceedings of the Third Conference on Machine Translation: Research Papers* (pp. 186–191). Association for Computational Linguistics.

Sułkowska, M. (2022). Phraseotranslation: Problems, Methods, Concepts. *Romanica Cracoviensia*, 1, 29–41. doi: 10.4467/20843917RC.22.003.15635

Tang, Y. et al. (2020). Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

Thompson B., & Koehn, P. (2019). Vecalign: Improved Sentence Alignment in Linear Time and Space. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp.1342–1348). Association for Computational Linguistics.

Acknowledgements

We extend our sincere thanks to Mrs. Anne-Lyse Minard, a specialist in Natural Language Processing (NLP) and Associate Professor at the Laboratoire Ligérien de Linguistique (LLL) at the University of Orléans, for her invaluable assistance in guiding us through computing techniques.

Contact information

Lian Chen

LLL, Université d'Orléans
LT2D, Cergy Paris Université
lian.chen@univ-orleans.fr

Wenjun Sun

L3i, Université de La Rochelle
wenjun.sun@univ-lr.fr

Flora Badin

CNRS-LLL
flora.badin@univ-orleans.fr

Enikő Héja, Kata Gábor, László Simon, and Veronika Lipp

GRAPH-BASED DETECTION OF HUNGARIAN ADJECTIVAL MEANING STRUCTURES VIA MONOLINGUAL STATIC EMBEDDINGS

Abstract The paper details the current state of an ongoing collaboration between Hungarian lexicographers and computational linguists. Our goal is to provide a comprehensive and consistent description of Hungarian adjectives, benefiting lexical semantics, lexicography and NLP. This thread of research focuses on identifying systematic semantic patterns of Hungarian adjectives and their typical subcategorization frames, with a particular emphasis on polysemous meanings. The proposed methodology is entirely unsupervised, reducing reliance on human intuition. It is based on a graph representation derived from adjectival static embeddings. The algorithm models adjectival semantic domains by specific subgraphs, namely, connected graph components. In the next step, potential subcategorization frames for the detected adjectival semantic domains, so called meaning structures, are also derived from corpus data. Then, a sample of the meaning structures is compared to the entries of the Concise Dictionary of Hungarian, evaluating the pros and cons of the proposed algorithm. Finally, as a further improvement, the automatically derived subcategorization frames were generalized.

Keywords automatic sense induction; monolingual lexicography; polysemy; unsupervised graph-based approach; adjectives

1. Introduction

1.1 Objectives – Lexicography, Lexical Semantics and NLP

This ongoing research aims to develop a consistent, data-driven method for describing Hungarian adjectives, which could benefit lexicography, lexical semantics, and NLP most importantly by providing adjectival sense distinctions. Here the sense distinctions are assigned to adjectival semantic domains rather than to specific adjectives, allowing for the identification of systematic semantic shifts. The data-driven nature empowers us to process a wide variety of adjectives, leading to a set of semantic features based on which more motivated sense distinctions can be made, thus resulting in a more consistent characterization of adjectival senses.

From a lexicographic perspective, the method aids in producing a consistent headword list by identifying common productive derivational processes in the formation of adjectives. This ensures that only those headwords not covered by explicit rules are included in the dictionary, which helps to maintain a consistent adjectival macrostructure.

The algorithm also supports the comprehensive and coherent lexical semantic description of Hungarian adjectives, which is currently lacking, although there are some lexical semantic investigations on Hungarian adjectives (Kiefer, 2008).

Moreover, by extracting explicit contextual clues – semantic features – for sense distinctions, our approach is expected to improve the quality of benchmark datasets to evaluate semantic knowledge in language models, particularly by increasing inter-annotator agreement in word sense disambiguation tasks (Pilehvar & Camacho-Collados, 2019).

1.2 The Scope of the Research

Although our previous investigation has shown that the proposed representation is able to filter homonym-candidates as well (cf. 2.3), the present research is focusing on *polysemous sense* distinctions. According to Haber & Poesio (2024, p. 3), polysemy is “far less well understood” than homonymy and “far less homogeneous than assumed in earlier literature.” To limit the scope of our investigation further, the present paper focuses on systematic meaning shifts, which operate on adjectival semantic domains rather than on a single adjective. These are sometimes referred to as idiosyncratic polysemy in the literature. The proposed algorithm models the definition of Apresjan (1974, p. 16) of regular polysemy. According to him, “Polysemy of a word *A* with the meaning a_i and a_j is called regular if [...] there exists at least one other word *B* with the meaning b_i and b_j , which are semantically distinguished from each other in exactly the same way as a_i and a_j and if a_i and b_i , a_j and b_j are nonsynonymous.”

bencés ‘Benedictine’	182	98.14	87.6
ciszterci ‘Cistercian’	49.29	26.43	331
cisztercita ‘Cistercian (fem)’	20.83	8.5	
dominikánus ‘Dominican’	32	11.83	
domonkosrendi ‘Dominican’	1.75		
ferences ‘Franciscan’	233.67	192.71	76.75
ferencrendi ‘Franciscan’	4.67	6.75	
jezsuita ‘Jesuit’	49.8	172.86	88.6
karmelita ‘Carmelite’	48.33	12.4	
piarista ‘Piarist’	49.67	94.6	193
premontrei ‘Premonstratensian’	36.88	20.17	73.5
	rendház ‘monastery’	szerzetes ‘monk’	iskola ‘school’
	monostor ‘monastery’	apát ‘abbott’	egyetem ‘university’
	kolostor ‘convent’	pap ‘priest’	gimnázium ‘high school’

Fig. 1: Part of the meaning structure of the lexical set ‘monastic orders’

esztelen 'senseless'	3.20	26.50	13.00	7.22
fékevesztett 'uncontrollable'	2.00	1.50	6.83	
fékezhetsen 'unstoppable'	2.33	2.00	6.62	
féktelen 'unrestrained'	21.60	12.00	7.50	21.64
oktalan 'unwise'	3.25	1.50	3.75	12.67
szertelen 'excessive'	6.00	2.00	9.88	
zabolátlan 'unbridled'	1.33	9.00		
	nevetés 'laughter'	költekezés 'spending'	rablás 'robbery'	vágy 'desire'
	száguldás 'speeding'	pazarlás 'wastefulness'	pusztítás 'destruction'	harag 'anger'
	zabálás 'binge eating'	ámokfutás 'rampage'		düh 'rage'
	ivászat 'drinking'			gyűlölet 'hatred'

Fig. 2: Part of the meaning structure of the near-synonyms of *féktelen* 'unrestrained'

As the automatically extracted matrices (Figure 1 and Figure 2) illustrate, according to this definition the adjectives in the semantic domains 'monastic orders' and 'unrestrained' equally exhibit regular polysemies characterized by the nominal sets at the end of each column, where each column corresponds to a specific sense candidate. For instance, let *A* be the word *bencés* 'Benedictine' and *B* be 'Jesuit'. As indicated by the nominal sets at the end of the columns in Figure 1, both 'Benedictine' and 'Jesuit' yield the senses of {'building', 'monastery', 'convent'} and {'clergy': 'monk', 'abbot', 'priest'} conforming to the criterion of regular polysemy as defined by Apresjan. However, we not only indicate whether a word *A* appears with one sense or not, but also assess the frequency of the given sense based on corpus data. These estimates appear in the cells of Figures 1. and 2. The algorithm is described in more detail in Section 2.

Following Atkins and Rundell (2008), we focus on two types of systematic meaning shifts: common word usage scenarios in lexical sets (semantic frames) and specific meaning shifts (systematic polysemy). Both types of meaning variations are important from a lexicographic perspective: lexical sets are defined by semantic frames linked to common sense knowledge (Figure 1), while near-synonyms are defined by systematic polysemies linked to lexical knowledge (Figure 2). The main advantage of our method is that it automatically extracts interpretable meaning patterns, ensuring consistent treatment of systematic sense variations across the vocabulary.

The research examines how effectively the proposed method aids in compiling the adjectival section of a monolingual Hungarian general purpose dictionary both at the levels of the macrostructure and the microstructure.

The paper is organised as follows: Section 2 introduces the workflow, Section 3 provides a brief overview on the Concise Dictionary of Hungarian, Section 4 evaluates the generated meaning structures from a lexicographic perspective, and Section 5 explores the possibility to generalize over the automatically attained adjectival subcategorization patterns.

2. Workflow

Our methodology extends the unsupervised graph-based approach described in Héja et al. (2022a; 2022b, and 2023), which was partially inspired by Ah-Pine and Jacquet (2009).

2.1 Extraction of Adjectival Semantic Domains From Corpus Data

In brief, adjectival meaning structure candidates were obtained from Hungarian monolingual data via the extraction of specific subgraphs from a graph of Hungarian adjectival word2vec embeddings.

Static embeddings, although more interpretable than LLMs, face criticism for potentially merging different senses (Camacho-Collados & Pilehvar, 2018). The conversion of embeddings into graphs is primarily aimed at addressing this issue of meaning conflation. Although the present paper focuses on polysemy, we claim that our approach is capable of filtering homonym-candidates as well, based on the presupposition that the various senses of a homonymous word form only accidentally coincide.

On the other hand, the graph representation has proven capable of capturing systematic semantic phenomena too, likely due to the tendency of words with similar usage scenarios or sense alternations to appear in more similar contexts compared to words with non-systematic sense relations.

This paper focuses on connected graph components, where any node can be reached from any other node by traversing edges. We found that connected components of the adjectival graph G correspond to clear-cut semantic domains.

2.2 Extraction of Connected Components From the Adjectival Graph

Let us briefly recap the technical details of the workflow:

1. 300-dimension static word2vec (Mikolov et al., 2013) representations were trained (Rehurek & Sojka, 2011) on the basis of 170 m sentences, part of the Hungarian Webcorpus 2.0 (Nemeskey, 2020).
2. An A adjacency matrix of the size $N \times N$ - N is the size of the adjectival vocabulary - was created where the $\langle i, j \rangle$ cell of A matrix was calculated using the usual cosine similarity between the vector representations of adj_i and adj_j .
3. The weighted undirected graph W was generated based on A adjacency matrix. Graph W is made up of N nodes and all nodes are interconnected in W with weights corresponding to the strength of semantic similarity between every pair of adjectives. Importantly, the induced graph's undirectedness is guaranteed by the symmetric nature of cosine similarity.

4. Subsequently, an unweighted graph G was created by binarizing W . A K cut-off parameter was used to eliminate edges with low strength. Each edge weight w was set to 1 if $w \geq K$ and w was set to 0 if $w < K$. As a result, graph G consists only of edges of the same strength ($w=1$); edges with $w = 0$ were omitted. Based on the results of manual evaluation K was set to 0.7 for the present experiments.
5. Finally, the connected graph components were extracted from graph G .
6. Below, we present two examples of the automatically extracted connected adjectival graph components: the first covers the lexical set of monastic orders, while the second contains adjectives from the semantic domain of *féktelen* ‘unrestrained’.

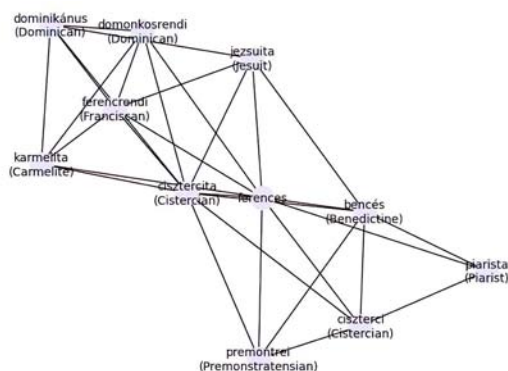


Fig. 3: Connected component of ‘monastic orders’

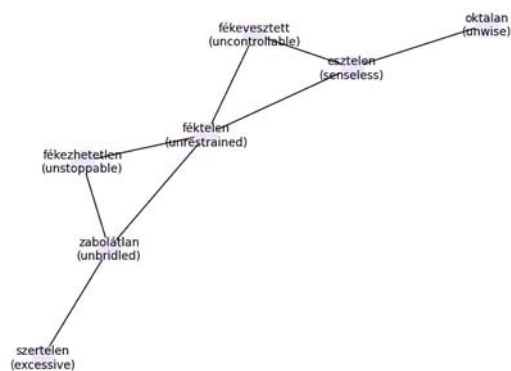


Fig. 4: Connected component for near-synonyms of ‘unrestrained’

2.3 Characterization of the Results

Using a $K=0.7$ cut-off parameter, the 53,490 input adjectives were divided into 1,750 connected graph components, each with at least 2 nodes, encompassing 11,474 adjectives. Of these, 1,748 components were entirely meaningful. Manual evaluation of components resulting from higher K cut-off values, and manual inspection of so called ego-graphs (a sub-graph made up of an ego-node and its adjacent nodes) confirmed our hypothesis, namely, that the two largest components (4505 and 669 adjectives) resulted from merging smaller meaningful components. The third largest component (355 adjectives) comprised personal names with derivational suffixes *-i* and *-s*. The fourth largest component (74 adjectives) included names derived from Hungarian cities with the *-i* suffix. Another notable component (43 adjectives) represented the semantic field of ages. Note that only 11,474 of the original 53,490 adjectives formed part of connected components with at least two nodes. This raises the question of what happened to the remaining cc. 40,000 adjectives. The parameter setting yielded 42,016 isolated nodes: these were deemed dissimilar to everything else by the algorithm. The manual evaluation of the 30 most frequent isolated adjectives showed that homonyms tend to end up as isolated nodes. This observation perfectly fits Apresjan’s definition of homonymy (Apresjan, 1974, p. 11), which describes

it as “a purely external coincidence of two or more words whose meanings have nothing in common.” Since a homonym has two (or more) unrelated meanings, the different randomly related sets of contexts in which it appears are combined during the training process. This results in a unique vector that is specific to the homonym itself, reflecting its various meanings.

The detailed investigation of the isolated nodes forms part of our future research. As Figure 5. illustrates, the connected component covering the semantic domain of country names does not contain *lett* (‘Latvian’, also meaning ‘became’), *észti* (‘Estonian’ and the accusative form of ‘wit’) and *ír* (‘Irish’ and ‘writes’), which ended up as isolated nodes.

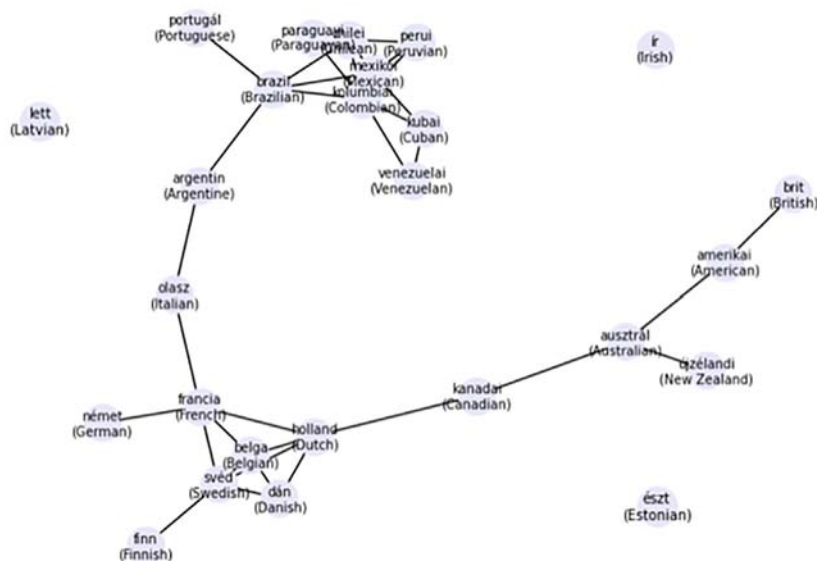


Fig. 5: The graph of country names with the three homonymic country names as isolated nodes

2.4 Creating the Meaning Structures

This study explores how automatically retrieved adjectival semantic domains can help characterise the meanings of adjectives, specifically identifying potential meaning structures of lexical sets or sets of near-synonyms (see Figures 1 and 2).

The workflow involves the following steps:

1. Manually select lexicographically interesting adjectival domains from the list of connected graph components (Figures 3 and 4).
2. Extract nouns that co-occur with any adjectives in the selected domains above a frequency threshold in the Hungarian National Corpus (Váradi, 2002) and cluster them using a hierarchical clustering algorithm (average linkage, cosine distance). The hierarchic clusters were flattened based on a cophenetic distance threshold within the clusters. The main advantage of hierarchical clustering is that the exact number of senses can vary, therefore,

do not have to be determined beforehand. The parameters were set based on the results of manual evaluation. The resulting nominal clusters serve as semantic feature candidates to characterise senses in the meaning structure of the adjectival semantic domains (columns in Figures 1 and 2).

3. Assess how well a nominal semantic category characterises a specific adjective by calculating the simple arithmetic mean of the number of adjective-noun co-occurrences for each nominal cluster (cells in Figures 1 and 2). This indicates the strength of association between each adjective and the corresponding semantic feature.
4. Automatically generate meaning structure matrices for the selected semantic domains (Figures 1 and 2).

3. The Concise Dictionary of Hungarian (CDH)

The Concise Dictionary of Hungarian (Pusztai, 2003) is a monolingual general-purpose dictionary containing 73,550 headwords, covering the vocabulary of both literary and colloquial Hungarian. To assess the extent to which the described method may contribute to the dictionary editing process, we first give a brief overview on CDH from multiple perspectives.

First, as for the adjectival vocabulary, the CDH comprises cc. 7,700 adjectival entries, comparable in size to that of the adjectival vocabulary assigned to connected components (11,474). However, the situation is not as clear as it may seem at first, since the part-of-speech categorization is not always consistent in CDH. In the case of adjectives the demarcation of nouns, adjectives and adjectival participles may raise some problems. We think that connected components can help in a more consistent PoS characterization, as due to the distributional similarity, adjectives belonging to the same connected component probably belong to the same PoS category/ies. For instance, although *megfelelő* ‘appropriate’ and *kellő* ‘adequate’ should both be considered adjectives, CDH assigns them falsely to different PoS-es.

The second problem is that although the CDH explicitly states (Pusztai, 2003, XII, XVII) that common derivational suffixes are enlisted and characterised as usual headwords, and thus, derivational forms are only enlisted if the suffixed form carries a new, distinct dictionary meaning compared to the base word, this principle is not adhered to consistently. At the same time, the semantic description of the derivational suffixes seems to be too loose and incomplete. Connected graph components offer a solution to provide a more explicit and clear-cut definition of derivational suffixes. In many cases, the extracted adjectives not only form tight semantic classes but also result from the same derivational process. This allows for an explicit and clear semantic description of the derivational process at both the input and output levels. For example, rather than using the vague and general definition of the derivational suffix -s in CDH as ‘<indicates that someone or something has someone or something, or that there is something in or on them>’, it would be much clearer to list specific cases of the derivational process describing precise and explicit semantic relations.

For instance, the semantic relations represented by the *-s* derivational suffix may differ in the contexts of <'course', 'meal'> and <'lane', 'road'>. This is evident in terms like *egyfogásos* 'one-course', *kétfogásos* 'two-course', or *háromfogásos* 'three-course' meals and *egysávos* 'one-lane' or *kétsávos* 'two-lane' roads.

The loose definition of derivational suffixes seems to cause problems at the level of the microstructure as well. For instance, although the derivational suffix *-s* is characterised as a headword, CDH comprises the transparently formed headword *kalapos* 'hatted': (adj) <*kalapot viselő* 'wearing a hat'>; <*aminek kalapja van* 'having a hat'>; <*amin kalapo(ka)t tartanak* 'on which hat(s) are kept'>. However, it is even more interesting that the third sub-sense of this meaning structure is basically non-existent in the light of synchronous corpus data.

Finally, the headword list is not consistent in the CDH either: for example, although in accordance with the edition principles the *-i* derivational suffix is neatly characterized as a headword, there are still cc. 1300 adjectival headwords with the *-i* suffix, several of them being completely transparent constructions.

4. Evaluating the Meaning Structures from a Lexicographic Point of View

During the evaluation phase, cc. 200 manually selected adjectives were examined representing various semantic relations, e.g., antonyms, synonyms, co-hyponyms. The automatically induced results were compared to the micro- and macrostructure of CDH.

4.1 Macrostructure

The analysis of the generated connected components revealed several further inconsistencies in the headword selection:

- Connected component I.: *csillagfényes* 'starlit', *csillagtalan* 'starless', *holdfényes* 'moonlit', *holdvilágos* 'moonlit', *teliholdas* 'full-mooned': neither *holdfényes* nor *teliholdas* were included as headwords in CDH.
- Connected component II.: *nullszaldós* 'break-even', *nyereséges* 'profitable', *rentábilis* 'profitable', *ráfizeteses* 'unprofitable', *veszteséges* 'loss-making': *ráfizeteses* was not included as a headword.
- Connected component III.: *gyakorlatlan* 'inexperienced', *gyakorlott* 'experienced', *rutinos* 'experienced', *rutintalan* 'inexperienced', *tapasztalatlan* 'inexperienced', *zöldfülű* 'rookie': *rutintalan* was not included as a headword.
- Connected component IV.: *bencés* 'Benedictine', *ciszterci* 'Cistercian', *cisztercita* 'Cistercian', *dominikánus* 'Dominican', *domonkosrendi* 'Dominican Order', *ferences* 'Franciscan', *ferencrendi* 'Franciscan Order', *jezsuita* 'Jesuit', *karmelita* 'Carmelite', *pannonhalmi* 'from Pannonhalma', *piarista* 'Piarist', *premontrei* 'Premonstratensian', *sarutlan* 'disalced': *domonkosrendi*,

pannonhalmi, and *ferencrendi* were not included as headwords. On the other hand, being only a collocate word, *sarutlan* should not be part of the connected component.

The method also works well with adjectives formed by productive derivation. For instance, adjectives with the derivational suffix *-nyi* form a semantically coherent adjectival class referring to distances, such as:

- Connected component V. *karnyújtásnyi* ‘within arm’s reach’, *kéznyújtásnyi* ‘within hand’s reach’, *kőhajításnyi* ‘within a stone’s throw’, *nyíllövésnyi* ‘within an arrow’s shot’, *puskalövésnyi* ‘within a gunshot’: based on Section 3, it is not obvious that these expressions should appear as headwords in CDH. But if one of them is included, every member of the component should be included, which is obviously not the case, as *kéznyújtásnyi* and *puskalövésnyi* are missing from CDH.

4.2 Microstructure

The method assists lexicographers by automatically finding the most important sub-senses, that is, meaning structures and the most frequent word combinations, which may serve as good examples in the dictionary entries. Referring back to Figure 1, we may say that the automatic methods reveal multiple subsenses (1. building, 2. position in the order, 3. educational institution).

Unfortunately, while the meaning structures demonstrate the potential to present these sub-senses coherently across the semantic domain, the senses are listed in the CDH inconsistently. For instance, when examining the entries for *sminkes* ‘makeup artist’, *pedikűrös* ‘pedicurist’ and *kozmetikus* ‘beautician’, we found that each is treated differently. *Sminkes* ‘Makeup artist’: I.(adj) <Related to or pertaining to makeup> II. (n) <A professional who performs makeup application in theatre or film studios.>; *Pedikűrös* ‘Pedicurist’ (n) <A foot care specialist.>; *Kozmetikus* ‘Beautician’ (adj and n) <A professional engaged in beauty care, particularly facial care.>

The PoS categories of all three entries should be uniform (both adjective and noun). Due to the strong semantic similarity, the definitions should also be standardised, for instance as follows: <A professional engaged in makeup application/foot care/facial care.>

Giving a closer look to another connected component comprising the adjectives *biszexuális* ‘bisexual’, *heteroszexuális* ‘heterosexual’, *homoszexuális* ‘homosexual’ and *leszbikus* ‘lesbian’, additional inconsistencies also were encountered in CDH: in three cases (*biszexuális*, *heteroszexuális* and *leszbikus*) the adjectival and nominal PoS categories are listed in a single definition, while in the remaining case (*homoszexuális*), the adjectival and nominal senses were told apart into separate definitions.

biszexuális 'bisexual'			13.2			
heteroszexuális 'heterosexual'		9	21.57	3		
homoszexuális 'homosexual'	14	16.4	54	22.67		49
leszbikus 'lesbian'	16	24.6	47.5		13	
	fesztivál 'festival'	szex 'sex'	anya 'mother'	megrontás 'molestation'	érintés 'touch'	propaganda 'propaganda'
	felvonulás 'parade'	házasság 'marriage'	katona 'soldier'	zaklatás 'harassment'		
		szerelem 'love'	férfi 'man'	erőszak 'violence'		

Fig. 6: Part of the microstructure of the lexical set 'homosexual'

Furthermore, Figure 6 also shows that the method yields a more nuanced structure of senses including: (1) person, specifically a man, a soldier; (2) relationship/affair; (3) abusive behaviour and (4) propaganda.

4.3 Pitfalls of Automatically Extracted Meaning Structures

A closer inspection of the CDH has shown that some of the headword adjectives are missing from graph G, which necessitates the re-generation of meaning structures based on a more balanced corpus. Another possible disadvantage of the automatically retrieved meaning structures is that they might be too fragmented. For instance, in the case of the connected component *bódult* 'dazed', *delíriumos* 'delirious', *ittas* 'drunk', *kábult* 'stunned', and *zavarodott* 'confused', the extracted elements of subcategorization frames 1. driver, motorcyclist, cyclist 2. car driver, chauffeur, police officer, vehicle driver 3. defendant 4. person, individual 5. girl, boy, woman, youth, man should be probably merged under the sense '<person>' in a general-purpose dictionary.

Below, we propose an approach that can help to find the right grain-size of semantic features to characterize senses.

5. Generalizing the Adjectival Subcategorization Patterns

The first results showed that the extraction of the nominal contexts suffers from two limitations: the generic parameter setting and the suboptimal clustering of the context noun groups. The generic parameter yields an uneven distribution of context nouns: some of them occur with many or most adjectives in the group, while others are very frequent with only one of the adjectives and are a likely candidate for a collocation. The suboptimal clustering on the other hand results in arbitrary changes in granularity: some noun groups might be too generic, while others are subdivided into several groups.

The problems above arise eventually from the fact that corpus co-occurrences were directly extracted. As a next step, we seek to construct a more abstract representation of context nouns that can (1) bring us closer to a semantic feature set to distinguish between senses instead of finding direct collocates, (2) overcome the idiosyncrasies (e.g., collocations) by generalizing over observed contexts, (3) ideally, selects the right level of generalization for semantic features. Thus, we examined the use of

huBERT, the Hungarian version (Nemeskey 2020) of BERT (Devlin & al., 2019) to represent adjectival subcategorization both individually and at the level of semantic fields defined by connected components. For each adjective, we construct a pattern in the form of “adjective [MASK]”.

We then extract the model’s predictions for the word following the adjective, i.e., the logit layer of the model output at the masked token position. It is a vector the size of the vocabulary, and – after softmax transformation – can be interpreted as the probability of each word in the vocabulary to appear in the given position. However, we use the values only for ranking. An advantage is that since we don’t extract a representation for the adjective itself, the method is insensitive to the tokenization of the adjective. From this vector we retrieve the first n elements of the vocabulary, those that are the most compatible with the preceding adjective. This will be the adjective’s subcategorization profile. For instance, as for the adjectives *antidemokratikus* ‘antidemocratic’, *autokrata* ‘autocratic’, *despotikus* ‘despotic’ and *diktatórikus* ‘dictatorial’, the obtained results show that the subcategorization profile of the semantic frame is composed of the nouns *rendszer* ‘system’, *hatalom* ‘power’, *politika* ‘politics’, *szervezet* ‘organization’ – obtained by selecting the nouns that appear in the subcategorization of all the adjectives in the group. The only outlier element of this group is ‘autocratic’ which according to the LLM subcategorises exclusively persons.

6. Conclusion and Future Work

The paper outlines the current state of an ongoing research, the ultimate goal of which is to provide a comprehensive and consistent description of the lexical semantics of Hungarian adjectives. Using a graph-based distributional approach, adjectival meaning structures were induced from monolingual corpus data, and compared to the Concise Hungarian Dictionary. While the results are promising from a lexicographic perspective, improvements are still needed: (1) The cut-off parameter K is currently based on manual evaluation and requires more mathematical grounding. (2) Over a third of the relevant adjectives are members of large connected components that merge various semantic fields, necessitating the dissection of the large connected components into more highly connected components. (3) Isolated nodes also may exhibit interesting semantic properties, most importantly homonymy and, therefore, need further investigation. (4) The coverage of results should be improved by retraining the algorithm on more balanced data. Finally, the potential for LLMs to complement the algorithm, particularly in generalizing subcategorization frames, should be further explored.

References

- Ah-Pine, J., & Jacquet, G. (2009). Clique-Based Clustering for Improving Named Entity Recognition Systems. In A. Lascarides et al. (Eds.), *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)* (pp. 51–59).
- Apresjan, J. D. (1974). Regular Polysemy. *Linguistics*, 12(142), 5–32.

Atkins, B. T. S., & Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford University Press.

Camacho-Collados, J., & Pilehvar, M. T. (2018). *From word to sense embeddings: A survey on vector representations of meaning*. <https://arxiv.org/abs/1805.04032>

Devlin, J. et al. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. <https://arxiv.org/abs/1810.04805>

Haber, J., & Poesio, M. (2024). Polysemy—Evidence from Linguistics, Behavioral Science, and Contextualized Language Models. *Computational Linguistics*, 50(1), 351–417. https://doi.org/10.1162/coli_a_00500

Héja, E., & Ligeti-Nagy, N. (2022a). A Clique-based Graphical Approach to Detect Interpretable Adjectival Senses in Hungarian. In D. Ustalov, Y. Gao, A. Panchenko, et al. (Eds.), *Proceedings of TextGraphs-16: Graph-based Methods for Natural Language Processing* (pp. 35–43).

Héja, E., & Ligeti-Nagy, N. (2022b). A proof-of-concept meaning discrimination experiment to compile a word-in-context dataset for adjectives – A graph-based distributional approach. *Acta Linguistica Academica*, 69(4), 521–548. <https://doi.org/10.1556/2062.2022.00579>

Héja, E., Ligeti-Nagy, N., Simon, L., & Lipp, V. (2023). An unsupervised approach to characterize the adjectival microstructure in a Hungarian monolingual explanatory dictionary. In Medved et al. (Eds.), *Proceedings of the eLex 2023 conference: Electronic lexicography in the 21st century: Invisible lexicography* (pp. 150–167). Lexical Computing.

Kiefer, F. (2008). A melléknevek szemantikája. In F. Kiefer (Ed.), *Strukturális magyar nyelvtan 4. A szótár szerkezete*. Akadémiai Kiadó.

Mikolov, T. et al. (2013). *Efficient estimation of word representations in vector space*. CoRR, abs/1301.3781.

Nemeskey, D. M. (2020). *Natural Language Processing Methods for Language Modeling*, [Ph.D. thesis]. Eötvös Loránd University.

Pilehvar, M. T., & Camacho-Collados, J. (2019). WiC: The word-in-context dataset for evaluating context-sensitive meaning representations. In Burnstein et al. (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1 (Long and Short Papers) (pp. 1267–1273).

Pusztai, F. (Ed.). (2003). *Magyar értelmező kéziszótár* [Concise Dictionary of Hungarian], Akadémiai Kiadó.

Rehurek, R., & Sojka, P. (2011). *Gensim—python framework for vector space modelling*. NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic, 3(2).

Váradi, T. (2002). The Hungarian National Corpus. In *Proceedings of the Second International Conference on Language Resources and Evaluation* (pp. 385–389).

Acknowledgements

This work was partially supported by the “PHC BALATON/HCB BALATON” programme (project number: 49855SF), funded by the French Ministry for Europe and Foreign Affairs, the French Ministry for Higher Education and Research and the Hungarian Ministry for Higher Education.

Our special thanks go to the reviewers for their valuable and insightful remarks.

Contact information

Enikő Héja

HUN-REN Hungarian Research Centre for Linguistics, Institute for Language Technologies and Applied Linguistics
heja.eniko@nytud.hun-ren.hu

Kata Gábor

INALCO, Équipe de Recherche Textes, Informatique, Multilinguisme (ERTIM)
kata.gabor@inalco.fr

László Simon

HUN-REN Hungarian Research Centre for Linguistics, Institute for Lexicology
simon.laszlo@nytud.hun-ren.hu

Veronika Lipp

HUN-REN Hungarian Research Centre for Linguistics, Institute for Lexicology
lipp.veronika@nytud.hun-ren.hu

Geraint Paul Rees and Isabel Gibert

A TEXTBOOK OR CHATGPT Which Helps Novice Programmers Most with Unknown Terms?

Abstract Existing research on ChatGPT in lexicography is undoubtedly valuable. However, it has tended to focus on metalexicographic concerns rather than effectiveness in resolving user queries directly. Moreover, it has mostly dealt with general-purpose English lexicography, often ignoring other languages and specific purposes. Focussing on 33 L1 Spanish users completing an introductory training course on the use of the Python programming language for linguistic research at a Spanish university, this study attempts to fill these gaps. Participants responded to ten multiple-choice questions designed to test understanding of basic programming terms. Approximately half received explanations from a respected introductory Python textbook written in Spanish. The remainder received ChatGPT 3.5-produced explanations written in Spanish. GPT-generated explanations offer performance advantages while textbook definitions offer advantages in processing time. In follow-up interviews, several participants reported feeling overwhelmed by the quantity of explanation provided by ChatGPT.

Keywords chatbots; ChatGPT; specific-purpose lexicography; Spanish; Python; user study

1. Introduction

Large language models paired with chatbots, such as ChatGPT, offer the potential of producing lexicographic resources, cheaply, quickly, and with minimal human involvement. Consequently, it is hardly surprising that since the launch of OpenAI ChatGPT in November 2022, there has been a substantial amount of research on its use for lexicography (c.f. de Schryver, 2023; Lew, 2024). Although undeniably valuable, there are several limitations to existing research. Firstly, much of this research has focused on lexicographers' and lexicography researchers' opinions of GPT's output rather than more direct measures of its effectiveness in resolving users' doubts about word meaning and use (e.g., Jakubiček & Rundell, 2023; Lew, 2023). Secondly, most studies have focused on the production of *general-purpose resources* for English learners (e.g., Rees & Lew, 2024). Ironically, users of these resources are perhaps amongst the best served by existing learners dictionaries and writing assistants. In comparison, the potential advantages offered by GPT in the creation of *specific-purpose lexicographic resources* for *less well-resourced languages* are greater. Thirdly, although much existing research has envisaged a place for GPT in the production of dictionaries, the possibility that users skip the dictionary altogether and go directly to the chatbot to resolve their language doubts has, for the most part, been only grudgingly acknowledged. Drawing a parallel with the shift from paper to online dictionaries, we believe that this is a distinct possibility, especially when one considers

the tendency observed in previous research for users to arrive at entry pages directly from search engines (Lorentzen & Theilgaard, 2012).

Concentrating on Spanish lexicography for specific purposes, in an attempt to address these gaps, this paper reports on a study involving post-graduate students at a Spanish university. These participants are L1 users of Spanish who were taking part in an online introductory training course on the use of the Python programming language for linguistic research. For L2 users of English like these novice programmers, the technical challenge of mastering new technology and the intellectual challenge of understanding new concepts are compounded by linguistic challenges. For example, although instructional material is available in Spanish, the meaning of key programming terms which were first coined in English is not always clear. Moreover, L2 English users are also at a disadvantage in understanding documentation and error messages which are often written in English (Becker, 2019; Guo, 2018). Students like these, then, often need to search for explanations of unknown programming terms in their L1. This study examines whether, in doing so, they would be best served by the kind of explanations found in a traditional programming textbook (Marzal & Gracia, 2020) or those generated by ChatGPT. We acknowledge, at the outset, that dictionaries may not be the most likely port of call for novice programmers wishing to clarify the meaning of an unknown term. However, we feel that the parallels between the comparison of information about an unknown term provided by a chatbot with that provided by a more traditional resource on the one hand, and the comparison of information provided in response to a specific language query by a chatbot with that provided by a dictionary, on the other, make this investigation worthwhile. With this in mind, we aim to answer the following research questions:

1. Do ChatGPT-generated explanations contribute to higher success than textbook explanations in a reception task dealing with key programming terms?
2. Do ChatGPT-generated explanations contribute to faster consultations than textbook explanations in a reception task dealing with key programming terms?

2. Methods

2.1 Participants and Setting

The study involved 33 post-graduate students and staff who were studying or teaching on the MA in the teaching of Spanish as a Foreign language or PhD programmes at the Hispanic Studies or Catalan departments at a university in Catalonia. These participants are L1 users of Spanish who were taking part in a voluntary introductory training course on the use of the Python programming language for linguistic research. The course took place online and comprised a two-hour video lecture. Before the class began, students were asked to complete a multiple-choice questionnaire aimed at testing their receptive knowledge of key programming terms. While completing the questionnaire, approximately half these students ($n = 16$) were

provided with definitions from a respected introductory Python textbook written in Spanish (Marzal & Gracia, 2020), while the remainder ($n = 17$) were provided with explanations produced by ChatGPT 3.5.

2.2 Instruments and Procedure

2.2.1 Multiple-Choice Questions and Explanations

In an attempt to ensure that the terms tested were likely to pose problems for novice Python programmers, we first created a pool of candidate questions inspired by existing online quizzes on basic programming terms. We focused on lexically oriented questions specifically intended to test knowledge of term meaning, leaving aside questions focussing on problem solving or analysis of code. In consultation with the course instructor and two other experienced programmers, we narrowed the pool of potential questions to ten items (see supplementary material). Happily, all the terms dealt with by these questions were present in the textbook from which the traditional explanations were drawn.

The questions were presented to participants using the 1KA online survey platform (1KA, 2023). After giving informed consent, participants were shown the first question along with four possible answers and a *no sé* ('I don't know') option. No explanation of the term of interest was provided at this stage. On clicking *no sé*, participants were given another opportunity to answer the question, this time with an explanation from the traditional textbook (TRAD) or one generated by GPT. For certain terms (e.g., *flotante* in supplementary material), the textbook contains something akin to a lexicographic definition. These were reproduced without their surrounding context. For other terms, the textbook provides more extensive explanation occasionally accompanied by code snippets (e.g., *def* in supplementary material). The GPT explanations were generated with ChatGPT 3.5. The prompt used was: "¿Qué significa X en el contexto de Python/programación?" ('What does X mean in the context of Python/Programming?'). The default settings for the free version of ChatGPT were used. The assumption being that this type of simple natural language prompt is likely to be used by novice programmers and that, like many software users and indeed online dictionary users (Trap-Jensen, 2010), these users are unlikely to alter the default settings. The code snippets from both sources were reproduced on the survey platform as image files.

2.2.2 Follow-up Interviews

In the weeks after the procedure, follow up interviews were carried out with six of the novice programmer participants to gain further insight about the relative merits of using the traditional resource or the more direct ChatGPT method to resolve language doubts. In these interviews, students were presented with both ChatGPT explanations and textbook definitions for an unknown term and asked to reflect on which, if any, they found more useful and why.

3. Results and Discussion

Two participants originally assigned to the GPT group are excluded from analysis as they abandoned the quiz without answering any items. The relatively large mean number (6.3 out of 10) of terms participants marked as unknown, suggests that the multiple-choice items were sufficiently challenging to probe the research questions and, more generally, lends weight to the assertion that understanding such terms poses a significant barrier to learning to code.

The following analysis focuses on items reported as being unknown. The first research question was answered by fitting a binary logistic mixed model which predicts success or failure for an item answered with the support of an explanation in the two conditions. The second research question was answered by fitting a linear regression model which predicts the time needed to answer an item in each condition. In both cases, model selection was carried out using the lme4 R package (Bates et al., 2015) with the assistance of the buildmer R package (Voeten, 2021). The R code for the fitting of both models can be found in the supplementary material.

3.1 Success

The final selected model for success included parameter estimates for random intercepts for items and participants. A summary of the best model is given in Figure 1.

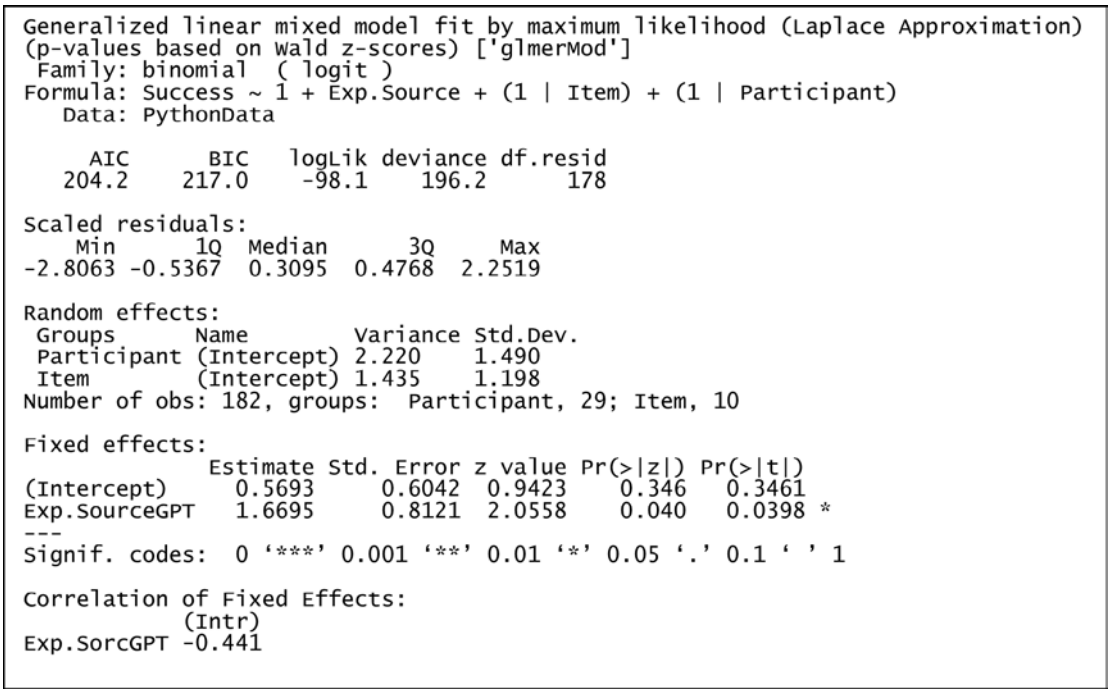


Fig. 1: A summary of the best logistic regression model

The parameter estimates for the GPT level are significantly higher ($z = 2.0558$, $p = 0.0398 < 0.05$) than those for the traditional textbook resource (TRAD) level. In Figure 2, these parameter estimates are displayed in terms of probabilities.

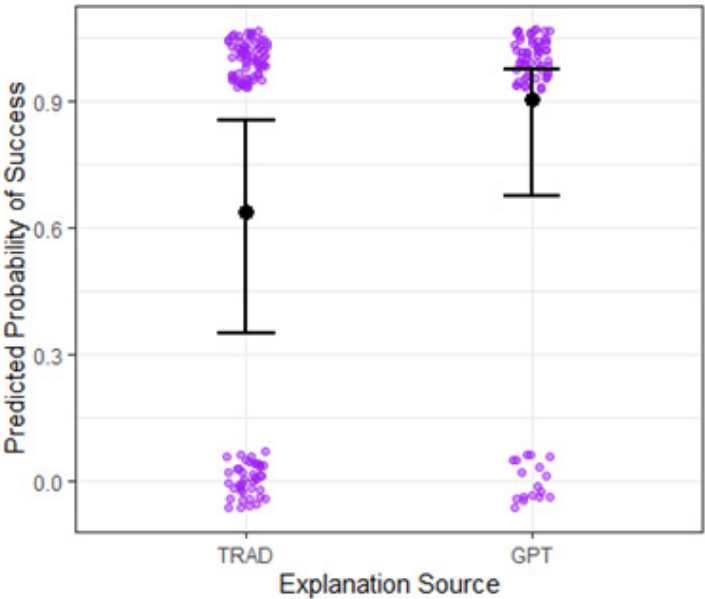


Fig. 2: Predicted probability of success for the two levels of explanation source, with 95% confidence intervals. Raw data points are shown jittered.

3.2 Time

Time measurements for every item were collected automatically using the paradata function of 1KA. Two clear cases of erroneous time measurements (416 & 2263 seconds)—possibly, due to participants leaving a question open and then returning to it later—are excluded from the analysis. The mean time per unknown item is markedly higher for GPT than TRAD (62.77 seconds vs. 35.22 seconds). Though the difference is exaggerated due to the skewness of the data. For further analysis, in line with accepted practice, we have transformed the data using logTime effectively compressing the scale of values, reducing the influence of extreme values thus making the distribution more symmetrical. A mixed regression model was fitted on logTime. The model parameters are given in Figure 3.

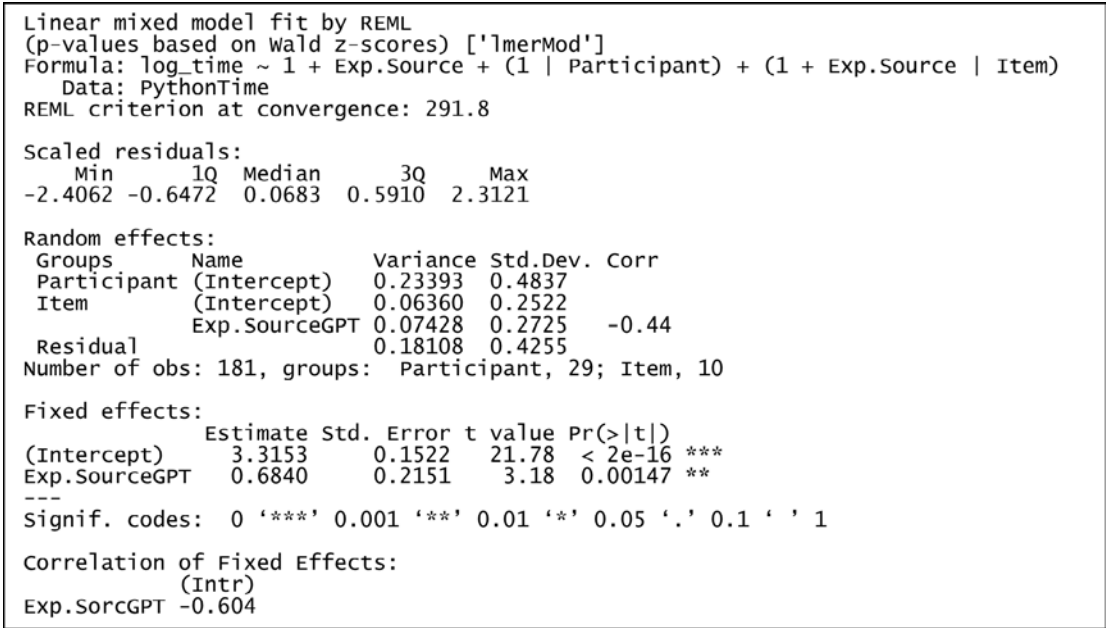


Fig. 3: A summary of the best linear regression model

Figure 4 gives the predicted time per item in seconds for the two explanation sources.

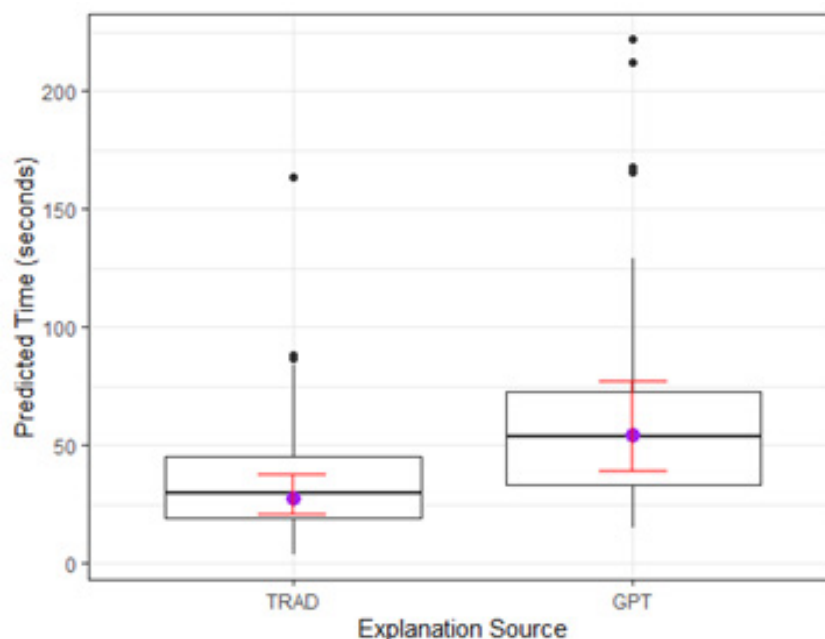


Fig. 4: Predicted time per item (in seconds) for the two levels of explanation source, with purple dots indicating mean and 95% confidence intervals are shown in red

3.3 Interviews

All six interviewees noted that the GPT definitions were more detailed. For certain items, they appreciated the extra information and examples. However, all but one interviewee reported feeling overwhelmed by the quantity of information provided at times (e.g., “ChatGPT gives a much larger and more complex explanation, which means that you forget a lot of information along the way”). Interviewees also expressed concerns about the appropriateness of the GPT explanations for novices (e.g., “ChatGPT offers a 250-word explanation, to put a number on it, of which 100 are incomprehensible. The examples don’t help either ‘the inverse of the square of 3’, no idea!”). More generally there were concerns about “irrelevant”, “repetitive”, and “redundant” information. These were echoed by the experienced programmers consulted when writing the multiple-choice questions. This chimes with debates in pedagogical lexicography about ‘good’ dictionary definitions and restricted defining vocabularies.

In line with previous research (Becker, 2019; Guo, 2018), four of the five interviewees who addressed the issue agreed that they, as users whose L1 is not English, face an additional linguistic challenge when learning to program. Not only do they have to comprehend key programming terms, they must also face the challenge of understanding new concepts (e.g., “Knowing English helps you to remind yourself about concepts and understand the commands”). See supplementary material for a (Spanish) summary of the interviews.

4. Conclusions

This study set out to address a gap in the research on the use of chatbots in non-English, specific-purpose, lexicographic tasks by examining a case similar to one in which the user could feasibly skip a lexicographic resource and resolve their queries directly with the chatbot. The success scores reported above suggest that ChatGPT can be more useful than a traditional textbook for this group of L1 Spanish novice programmers in supporting comprehension of unknown programming terms. Taken in isolation, this suggests that users might skip the traditional reference resource altogether and resolve their linguistic doubt directly using ChatGPT. However, the significantly greater mean time needed to resolve queries using ChatGPT and participants' reservations about the relevance and quantity of information provided by ChatGPT suggest that there may be a role for more carefully curated reference resources in conjunction with chatbot queries.

Care should be taken not to overgeneralise these conclusions. They were based on the responses of a relatively small number of participants to questions designed to test receptive knowledge of only ten basic programming terms. It is also notable that while this Spanish study counters English dominance in AI research, there are a great number of languages which are even less researched. Furthermore, the data collection instrument underplays the real-world affordances of ChatGPT. For example, on stating that they did not know a term, users were immediately presented with explanations from one of the sources. This meant that neither the time needed to find an item in a traditional resource and other accessibility advantages (Rees, 2023), nor the possibility of asking follow-up questions in ChatGPT were accounted for. Finally, chatbot technology is constantly evolving, not only in terms of underlying models, but also in terms of closer integration with development environments and productivity software (e.g., BethanyJep, n.d.) in a similar way to previous integration of dictionaries with word processors via writing assistants (e.g., Frankenberg-Garcia et al., 2019). These limitations present interesting avenues for future research.

References

- 1KA. (2023). [Computer software]. Faculty of Social Sciences. <https://www.1ka.si>
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1–48. <https://doi.org/10.18637/jss.v067.i01>
- Becker, B. A. (2019). Parlez-vous Java? Bonjour La Monde!= Hello World: Barriers to Programming Language Acquisition for Non-Native English Speakers. In M. Marasoiu, L. Church, & L. Marshall (Eds.), *Proceedings of the 30th Annual Workshop of the Psychology of Programming Interest Group, PPIG 2019, Newcastle University, UK, August 28–30, 2019*. Psychology of Programming Interest Group. <https://ppig.org/papers/2019-ppig-30th-becker/>
- BethanyJep. (n.d.). *Using GitHub Copilot with Python—Training*. Retrieved 29 May, 2024, from <https://learn.microsoft.com/en-us/training/modules/introduction-copilot-python/>

de Schryver, G.-M. (2023). Generative AI and Lexicography: The Current State of the Art Using ChatGPT. *International Journal of Lexicography*, 36(4), 355–387. <https://doi.org/10.1093/ijl/ecad021>

Frankenberg-Garcia, A., Lew, R., Roberts, J. C., Rees, G. P., & Sharma, N. (2019). Developing a writing assistant to help EAP writers with collocations in real time. *ReCALL*, 31(1), 23–39. <https://doi.org/10.1017/S0958344018000150>

Guo, P. J. (2018). Non-Native English Speakers Learning Computer Programming: Barriers, Desires, and Design Opportunities. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–14. <https://doi.org/10.1145/3173574.3173970>

Jakubiček, M., & Rundell, M. (2023). The End of Lexicography? Can ChatGPT Outperform Current Tools for Post-editing Lexicography? In M. Medveď, M. Měchura, I. Kosem, J. Kallas, C. Tiberius, & M. Jakubiček (Eds.), *Electronic lexicography in the 21st century (eLex 2023): Invisible Lexicography. Proceedings of the eLex 2023 conference* (pp. 518–533). Lexical Computing CZ s.r.o.

Lew, R. (2023). ChatGPT as a COBUILD lexicographer. *Humanities and Social Sciences Communications*, 10(1), Article 1. <https://doi.org/10.1057/s41599-023-02119-6>

Lew, R. (2024). Dictionaries and lexicography in the AI era. *Humanities and Social Sciences Communications*, 11(1), 1–8. <https://doi.org/10.1057/s41599-024-02889-7>

Lorentzen, H., & Theilgaard, L. (2012). Online dictionaries – how do users find them and what do they do once they have? In R. V. Fjeld, & J. M. Torjusen (Eds.), *Proceedings of the 15th EURALEX International Congress* (pp. 654–660). Department of Linguistics and Scandinavian Studies, University of Oslo.

Marzal, A., & Gracia, I. (2020). *Introducción a la programación con Python*. Publicacions de la Universitat Jaume I.

Rees, G. P. (2023). Online Dictionaries and Accessibility for People with Visual Impairments. *International Journal of Lexicography*, 36(2), 107–132. <https://doi.org/10.1093/ijl/ecac021>

Rees, G. P., & Lew, R. (2024). The Effectiveness of OpenAI GPT-Generated Definitions Versus Definitions from an English Learners' Dictionary in a Lexically Orientated Reading Task. *International Journal of Lexicography*, 37(1), 50–74. <https://doi.org/10.1093/ijl/ecad030>

Trap-Jensen, L. (2010). One, Two, Many: Customization and User Profiles in Internet Dictionaries. In A. Dykstra, & T. Schoonheim (Eds.), *Proceedings of the 14th EURALEX International Congress* (pp. 1133–1143). Fryske Akademy.

Voeten, C. C. (2021). *Using 'buildmer' to automatically find & compare maximal (mixed) models*. <https://cran.r-project.org/web/packages/buildmer/vignettes/buildmer.html>

Supplementary Material

Supplementary material including the R code for the model selection and figures, the multiple-choice questions and explanations, the participant response data, and a Spanish summary of the interviews can be found here: <https://osf.io/9rbdw/>

Acknowledgements

The authors are grateful for the support of the Research Group in Language and Technology (ReLaTe) (2021 SGR 00151).

Contact information

Geraint Paul Rees

Department of Translation and Language Sciences, Universitat Pompeu Fabra, Barcelona, Spain
geraintpaul.rees@upf.edu

Isabel Gibert

Department of Romance Studies, Universitat Rovira i Virgili, Tarragona, Spain
isabel.gibert@urv.cat

Bálint Sass

THE “DEPENDENCY TREE FRAGMENTS” MODEL FOR QUERYING A CONSTRUCTICON

Abstract In this paper, we argue that it is beneficial for a constructicon to have a query interface where the user can enter arbitrary text. To allow this, we present a dependency-tree-based model for representing constructions, and show that this model can serve as a basis for working out a user-friendly query interface which analyses the free-text user query and matches the constructicon entries to it, revealing all constructions from the query text for the user, without expecting any knowledge of construction grammar.

Keywords constructicon; construction; dependency tree; dependency tree fragments; Universal Dependencies; Hungarian

1. Motivation: The User’s Perspective

We consider inventories of constructions (i.e., constructicons) as a replacement for dictionaries, not as a supplement to them. We agree with Hilpert (2014, p. 2) as he states that “a person’s knowledge of language consists of nothing but constructions”. Also agreeing with Goldberg (2006) who includes bound and free morphemes as well into the set of cxns¹ and contradicting Haspelmath (2023) who excludes items without open slots, we consider every linguistic form having a meaning a cxn regardless of how complicated or how simple it is. This way, a ccn covers everything a dictionary does and much more.

The basic way of interaction with a dictionary is to provide an input word and obtain its dictionary entry. In the past, you were not able to do this without knowing some kind of canonical form of the word. Traditional dictionaries have had cross-referencing to help finding the canonical form corresponding to an irregular form, while modern dictionaries do this for all regular forms as well (look up e.g., *books* in OALD (Oxford University Press, 2023)). We think that a ccn should have a similar functionality extended to possible multiword inputs: it should be able to return the (canonical form of) cxns present in the input text regardless of the specific form they have.

As a kind of dictionary, a ccn is intended for the public: mainly for language learners and all those who are interested. We do not want it to be “probably best suited for users with better than average knowledge of linguistics” (cf. Lyngfelt et al., 2018b, p. 92). We do not expect the user to know some canonical form of the cxn in question not even the theoretical notion of what a cxn is. Note that the canonical form of cxns may be much more complex than a canonical form of words. It is some kind of machine-readable formal representation handling different kinds of morphemes, connections, slots and fillers.

¹ To improve readability, we use the abbreviation *ccn* for constructicon and *cxn* for construction.

We expect the user to have a (multiword) linguistic utterance in mind, and that he/she wants to uncover its structure and meaning. We think that the solution is the following. To function similarly to a dictionary a ccn should have a dynamic machinery in the background: it should be able to accept arbitrary short text, analyse it on the fly and reveal the cxns in it to the user. For example, it should reveal the cxn ‘*take something for granted*’ for inputs like *I take it for granted*, or even *some French journalist actually took this story for granted*, or *it was taken for granted*.

Our proposed solution can be compared to the following approaches: (1) to present a browsable list of cxns to the user; (2) to work out a sophisticated query interface which reflects the subtleties of the cxn-representation; (3) to give the user a formal query language which works on the cxn-representation. Ccns tend to use the first two approaches (cf. the Swedish ccn (Lyngfelt et al., 2018b) and the Russian ccn (Janda et al., 2020; Bast et al., 2021)). None of these three approaches are ideal in our opinion. The first approach is too limiting in two ways: firstly, it is something like browsing pre-created answers instead of having the possibility to ask a question, secondly, allowing to look at just one cxn at a time it does not make it possible to investigate the interconnections of cxns. The other two approaches expect some knowledge of the cxn-representation which we would like to hide from the user as we argued above.

To work out our approach we need (1) a representation which is capable to represent any kind of cxns; (2) to store the entries of the ccn in this form; (3) an algorithm to analyse the input text according to the representation; and (4) another algorithm to match the ccn-entries to the representation of the input query. To elaborate this representation, the starting idea is to use dependency trees with filled vertices and empty vertices (corresponding to open slots) as well to represent cxns.

2. The Hungarian Constructicon

We are implementing the above approach for the Hungarian ccn which is in preparation.

The Hungarian Constructicon project (Sass, 2023) follows the approach of processing an existing dictionary and extracting cxns from it to create the database of the ccn. Not just headwords become cxns, but also expressions that are hidden in the “expressions” part and even in the “examples” part of the entries. All of these become ccn-entries (or head-constructions (hcxns)) of their own, together with their definitions. For example, Hungarian counterparts of *action* and *take part in* will be hcxns of the same right. After the above processing, what is listed as hcxn in the ccn is accepted as a cxn. In other words, we leave the burden of defining cxns to the original dictionary.

3. The “Dependency Tree Fragments” Model

Based on the first section, we can consider a concrete single morpheme (Goldberg, 2006, p. 5) a cxn, and at the other end of the scale we can consider an abstract

grammatical rule (a constructional schema (Diessel, 2023, p. 16)) a cxn as well. Cxns consist of two kinds of elements: fixed elements (fillers) and open slots where other cxns may be inserted as a filler (Lyngfelt et al., 2018a). Taking the above examples, a single morpheme is a fixed element in itself, and a grammatical rule contains only open slots. The main advantage of the constructionist approach is precisely that it can handle all of the above in a unified framework as cxns.

Most interesting cxns are in between, these are the cxns that have both kinds of elements: for example, *take part in*. Beyond its three fixed elements, this cxn has two open slots as well, one for the subject and one for the *in*-PP. This cxn is complete only together with all of these, it should be included in a ccn as an entry in its complete form. We can think of these mixed expressions as the genuine building blocks of language, of which, words and grammatical rules are only special cases.

Building on the concept of “dependency tree fragments” (e.g., Morimoto et al., 2016), our basic idea is the following: dependency-tree-like structures can represent not just texts but cxns as well. The main difference to handle is this: while in texts all slots are filled with some words, in cxns some slots are not filled. The essential point is that we can represent open slots appropriately, specifically, by edges having no vertex at the bottom end of the edge (Figure 1).

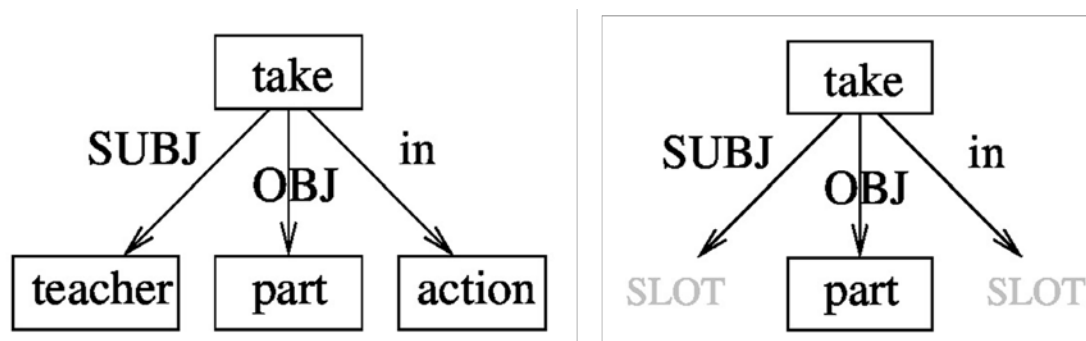


Fig. 1: Left side: representation of text *the teacher takes part in the action*. Right side: representation of cxn *take part in* together with its two open slots. Open slots are represented by edges having no vertex at the bottom end.

As an alternative to drawing trees (cf. Figure 1) we can use the following equivalent textual representation respectively:

[take /SUBJ:teacher /OBJ:part /in:action]

[take /SUBJ:_ /OBJ:part /in:_]

An edge is marked by [/SLOT], a vertex is marked by [:filler], and the notation for an open slot is [:_] (underscore).

As our implementation is for Hungarian, we will use Hungarian examples in the remaining part of this paper.

4. The Same Analysis for Constructicon Entries and Query Texts

In order to be able to compare the free-text queries with our ccn-entries we use the same representation formalism – namely the “dependency tree fragments” model described in Section 3 – for both of them.

For creating the representation, we utilize a classical natural language processing pipeline. For tokenization, morphological analysis and POS-tagging, we use the *emtsv* system (Indig et al., 2019; Mittelholcz, 2017; Novák et al., 2016; Orosz & Novák, 2013) combined with the dependency-analysis module of the *udpipe* system (Straka et al., 2016) which is integrated into *emtsv*, and provides an output which meets the Universal Dependencies standard.

Free-text queries are obviously texts. Initially, ccn-entries has a simple textual form as well, as they come from a dictionary. The important difference that hcxns have open slots, is somewhat hidden in this form in an unformalized way. A ccn-entry can be the following, for example: *‘részt vesz valamiben’* (‘take part in something’). What we want to emphasize here is what appears as *‘valami’* (‘something’) in the textual form of a hcxn is in fact an open slot in the vast majority of cases. Our representation allows us to formalize this information, so after running the analysis we transform it to an open slot on the fly. It is important that although *‘valami’* is deleted, its suffix *‘ben’* is retained. This is a Hungarian inessive (INE) case marker with a meaning similar to the English preposition *in*. Clearly, it should be kept, as it marks the open slot in the cxn in question. This is how the ccn-entry *‘részt vesz valamiben’* becomes [vesz /OBJ:rész /INE:valami] after analysis and then [vesz /OBJ:rész /INE:_] after the open slot transformation. The latter form is stored in the ccn database alongside the original textual form of the ccn-entry.

Concerning slots around verbs in Universal Dependencies, beyond *nsubj* (subject), *obj* (object) and *iobj* (indirect object), we have *nmod:obl* (oblique) among the most frequent ones. The latter category is too coarse-grained for our purposes. We decided to retain the case of the oblique as well, in other words we break down the oblique category by case, because the case carries essential information about the slot.

Having run the analysis on all ccn-entries we obtain a ccn augmented with a formal representation for each hcxn. Also, we have the same algorithm at hand to provide formal representation for free-text input queries.

5. An Algorithm for Identification of cxns in Query Text

Our goal is to extract hcxns from free-text user queries. Fixed, continuous cxns are easier to handle by some string-matching-based methods, but Hungarian verbal cxns pose a more difficult task, as they can be non-continuous and can be in different word orders. The proposed representation helps us by abstracting from these properties.

The main algorithm of our proposed system is intended for identifying cxns in free-text queries. The big picture is the following: (1) the user enters a query; (2) the system assigns “dependency tree fragments” representation to it by automatic linguistic analysis; (3) then compares the query against the ccn-entries in our ccn database which have the same representation; (4) recognizes cxns in the query; (5) and presents them to the user finally.

A query can contain several cxns. In the sentence ‘*a tanár részt vesz az akcióban*’ (‘the teacher takes part in the action’) the main cxn is clearly ‘ $SLOT_{SUBJ}$ részt vesz $SLOT_{INE}$ ’ (‘ $SLOT_{SUBJ}$ take part $SLOT_{in}$ ’). But as words are cxns as well, there are others: ‘*tanár*’ (‘teacher’) and ‘*akció*’ (‘action’) not to mention the two article-cxns. We would like to show all of them to the user.

Our most important observation is the following. Notice that all cxns are a specific part of the representation tree of the query, and on the other hand, these specific parts add up to the tree without remaining fragments. So our task is to “split up” the tree into cxns somehow. An illustration can be seen in Figure 2.

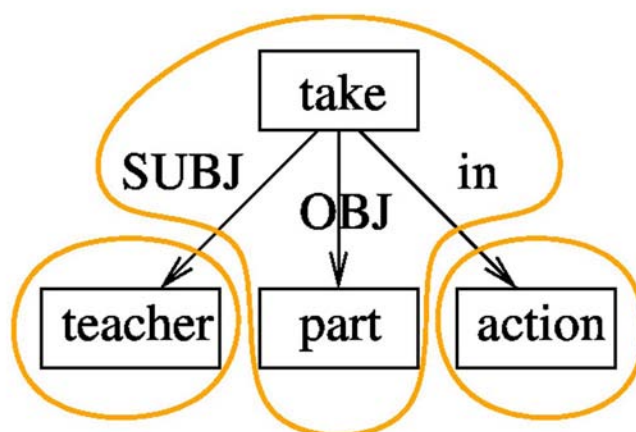


Fig. 2: The three cxns into which the sentence ‘*a tanár részt vesz az akcióban*’ (‘the teacher takes part in the action’) is split up by our “dependency tree fragments” algorithm. The two article-cxns are not depicted.

We propose the following algorithm.

1. start from the root vertex of the tree of the query text;
2. taking the ccn, search for hcxns which *matches* here;
3. take the longest matching hcxn if there are several;
4. *remove* the found hcxn from the tree;
5. after the removal the tree falls apart into smaller trees;
6. take these smaller trees one by one and start over the algorithm.

In step 2 above, “ H (a hcxn) *matches* T (query text)” means that the root is the same in their representation, T contains all the direct edges branching from the root vertex what H contains, and the vertices at the end of these edges are the same or H has an open slot there. In other words, an open slot in the hcxn matches any fillers present

in the query. If H has further vertices and edges, they should match similarly in a recursive manner. Basically, this is something like determining whether H is in a subset-like relation with T .

Performing *remove* (in step 4 above), if the cxn to remove has filled slots at its lowest level, we keep these vertices in order to being able to retain the edges starting from them.

If we get to a single vertex/filler (e.g., ‘*tanár*’ (‘teacher’) in Figure 2), we apply morphological analysis to split up the word into a series of morphemes. Then we attempt to match the entries of the ccn onto this series. This is a similar procedure what we do with trees, but much simpler, as the morpheme structure of a word is not a tree but a simple chain.

The reader can follow the operation of the algorithm in Figure 2. The first step of splitting up the sentence into cxns is matching the hcxn on the right side of Figure 1 onto the original text. After removing this hcxn, what remains is two one-vertex tree, each of them representing a monomorphemic filler.

6. Examples

The example we have analysed throughout the paper in Figures 1 and 2 had the following structure: at root there was a multi-element cxn with two slots, and the fillers in both of the slots were simple single-element fillers. Of course, the algorithm works well on inputs of a different structure as well. We illustrate this through two more examples.

A single-element root hcxn can have a multi-element filler as we see in the example ‘*megérkezik az alapító tag*’ (‘the founding member arrives’). The representation of the query is [megérkezik /SUBJ:tag /SUBJ/amod-att:alapító /SUBJ/det:az]. We note that [/V1/V2:F] notation means a two-edge path in the tree from the root: a [V1] vertex, then a [V2] vertex connected to the end of [V1], and [V2]’s own filler is [F]. After processing the root element as a single word we get the following representation: [tag /amod-att:alapító /det:az]. We have ‘*alapító tag*’ (‘founding member’) as a hcxn with the representation [tag /amod-att:alapító] in the ccn. We can see that this will match because the query fragment being processed contains all the edges that it has, and there is no contradiction between fillers. At the end we will have three cxns: ‘*megérkezik*’ (‘arrive’), ‘*alapító tag*’ (‘founding member’) and an article-cxn.

Of course there can be more than one multi-element hcxn in the query text. In the case of ‘*az alapító tag részt vesz az akcióban*’ (‘the founding member takes part in the action’) not just the root element, but one of the fillers is a multi-element hcxn. The algorithm handles this similarly to the above. As we see, cxns occur as fragments of dependency trees as the method’s name suggests.

It is worth to mention that the algorithm implements a kind of “fallback” mechanism. First at the level of words, then at the level of morphemes. If no multi-element hcxns

are found at the current root vertex, then the algorithm falls back to process the vertices (fillers, words) one by one; and then if no single-element hcxn is found at the current root vertex, the filler’s morphemes are processed one by one. As an example for the latter, ‘*eseményekről*’ (‘about events’) consists of three morphemes which are three separate cxns: ‘*esemény*’ (‘event’), ‘*-ek*’ (plural marker) and ‘*-ről*’ (‘about’).

7. Demo Implementation

The main contribution of this paper is the model itself together with the two algorithms operating it described above. In order to give the reader an impression of how this might work in practice we created a demo implementation as well. It has a limited feature set, its main purpose is to demonstrate the essential properties of the model, the basic operation and feasibility of the proposed system and show its potentials. To create a fully operational system is a long term goal.

The “dependency tree fragments” demo is available here as part of the Hungarian Constructicon project: <https://ccn.nytud.hu> (username: **demo**, password: **letssee**). You can see how the system works for examples mentioned in the paper, and also, you can test the system entering other similar short Hungarian texts.

References

- Bast, R., Endresen, A., Janda, L. A., Lund, M., Lyashevskaya, O., Mordashova, D., Nessel, T., Rakhilina, E., Tyers, F. M., & Zhukova, V. (2021). The Russian Constructicon. An electronic database of the Russian grammatical constructions. <https://constructicon.github.io/russian>
- Diessel, H. (2023). *The Constructicon*. Cambridge University Press.
- Goldberg, A. E. (2006). *Constructions at work: The nature of generalization in language*. Oxford University Press.
- Haspelmath, M. (2023). On what a construction is. *Constructions*, 15(1), 1–15. doi 10.24338/cons-539
- Hilpert, M. (2014). *Construction Grammar and Its Application to English*. Edinburgh University Press.
- Indig, B., Sass, B., Simon, E., Mittelholcz, I., Vadász, N., & Makrai, M. (2019). One format to rule them all – The emtsv pipeline for Hungarian. In A. Friedrich, D. Zeyrek, & J. Hoek (Eds.), *Proceedings of the 13th Linguistic Annotation Workshop* (pp. 155–165). Association for Computational Linguistics.
- Janda, L., Endresen, A., Zhukova, V., Mordashova, D., & Rakhilina, E. (2020). How to build a constructicon in five years: The Russian example. *Belgian Journal of Linguistics*, 34, 162–175.
- Lyngfelt, B., Borin, L., Ohara, K., & Torrent, T. T. (Eds.). (2018a). *Constructicography: Constructicon development across languages*. John Benjamins.

Lyngfelt, B., Bäckström, L., Borin, L., Ehrlemark, A., & Rydstedt, R. (2018b). Constructicography at work: Theory meets practice in the Swedish Constructicon. In B. Lyngfelt, L. Borin, K. Ohara, & T. Torrent (Eds.), *Constructicography: Constructicon development across languages* (pp. 41–106). John Benjamins.

Mittelholcz, I. (2017). emToken: Unicode-képes tokenizáló magyar nyelvre. [emToken: a Unicode-capable tokenizer for Hungarian.]. In V. Vincze (Ed.), *Proceedings of MSZNY2017* (pp. 70–78). Szegedi Tudományegyetem, Informatikai Tanszékcsoport.

Morimoto, A., Yoshimoto, A., Kato, A., Shindo, H., & Matsumoto, Y. (2016). Identification of Flexible Multiword Expressions with the Help of Dependency Structure Annotation. In E. Hajičová, & I. Boguslavsky (Eds.), *Proceedings of the Workshop on Grammar and Lexicon: interactions and interfaces (GramLex)* (pp. 102–109). The COLING 2016 Organizing Committee.

Novák, A., Siklósi, B., & Oravecz, Cs. (2016). A New Integrated Open-source Morphological Analyzer for Hungarian. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odiijk, & S. Piperidis (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA).

Orosz, G., & Novák, A. (2013). PurePos 2.0: a hybrid tool for morphological disambiguation. In R. Mitkov, G. Angelova, & K. Bontcheva (Eds.), *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2013)* (pp. 539–545). INCOMA Ltd. Shoumen.

Oxford University Press (2023). Oxford Advanced Learner's Dictionary. <https://www.oxfordlearnersdictionaries.com>

Sass, B. (2023). From a dictionary towards the Hungarian Constructicon. In M. Medveď, M. Měchura, C. Tiberius, I. Kosem, J. Kallas, M. Jakubíček, & S. Krek (Eds.), *Electronic lexicography in the 21st century (eLex 2023): Invisible Lexicography. Proceedings of the eLex 2023 conference* (pp. 534–544). Lexical Computing CZ s.r.o.

Straka, M., Hajič, J., & Straková, J. (2016). UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odiijk, & S. Piperidis (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 4290–4297). European Language Resources Association (ELRA).

Acknowledgements

The research described in this paper was supported by the OTKA (K 147452) grant of the National Research, Development and Innovation Office (NKFIH).

Contact information

Bálint Sass

HUN-REN Hungarian Research Centre for Linguistics, Institute for Lexicology
sass.balint@nytud.hu

Dragana Špica and Benedikt Perak

ENHANCING JAPANESE LEXICAL NETWORKS USING LARGE LANGUAGE MODELS

Extracting Synonyms and Antonyms with GPT-4o

Abstract This study presents an innovative approach to crafting and enhancing Japanese lexical networks by incorporating large language models (LLMs), especially GPT-4o, utilizing data from Matsushita's (2011) Vocabulary Database for Reading Japanese to accommodate various proficiency levels. Through this process, we extracted a total of 137,870 synonym relations and 54,324 antonym relations, forming a network comprising 104,427 nodes. A portion of the dataset underwent manual evaluation to determine the accuracy of the extracted synonym relationships, yielding an average evaluation score of 4.08 out of 5. Our findings demonstrate that almost 20% of extracted nouns are (near) synonyms, while the rest have various relation types to the source word including hyponymy, hypernymy, meronymy, class membership etc. The study emphasizes the synergy between AI-driven data generation and traditional lexicographic expertise, offering a scalable and adaptable framework for diverse linguistic applications, with implications for computational linguistics and NLP technologies.

Keywords Japanese lexical networks; large language models; GPT synonyms and antonyms; Semantic graph analysis; AI-enhanced lexicography

1. Introduction

Lexical networks, graph representations of semantic and syntactic relationships between words, have proven invaluable in various natural language processing (NLP) tasks. These networks, when accurately constructed, offer insights into language structure, facilitate tasks like word sense disambiguation, and support language education (Navigli, 2009). Traditionally, building such networks has been a labor-intensive process, relying on expert lexicographers and time-consuming manual annotation.

Recent advancements in large language models (LLMs) have opened new avenues for automating and scaling the creation of lexical resources. These models, trained on vast amounts of text data, have demonstrated remarkable abilities in generating text, translating languages, and understanding complex instructions. However, their potential for constructing lexical networks remains largely unexplored.

This study aims to bridge this gap by proposing a novel methodology for constructing and enriching Japanese lexical networks using GPT-4o. We leverage the model's capacity to generate synonyms and antonyms, coupled with graph-based techniques, to create a comprehensive and interconnected representation of the Japanese lexicon. By incorporating additional linguistic features and evaluating the network's quality,

we seek to demonstrate the effectiveness of our approach and its potential for broader applications in NLP and language education.

We introduce a methodology for extracting and examining lexical relationships, emphasizing the derivation of associative concepts through synonym and antonym connections, as well as semantic class labeling in the Japanese language. This approach utilizes prompts to query a large language model about word relationships, subsequently aggregating the responses to construct a comprehensive semantic network.

The integration of LLMs and graph procedures in the study of lexical relationships offers a modern approach to understanding and organizing vocabulary, particularly in the context of Japanese language learning and linguistic research (Aotani & Takahashi, 2021). The resulting lexical network is a kind of abstraction of the mental repository of word knowledge that facilitates efficient information retrieval and processing during language comprehension and production (Kovács et al., 2021).

State-of-the-art Large Language Models (LLMs) are pretrained on extensive text data from multiple languages, with the vast majority sourced from publicly available Internet resources (e.g., the latest Common Crawl dataset, which includes data from over 3 billion pages). Among this data, the Japanese corpus constitutes 0.11% of the total training dataset for GPT3 (Kawahara et al. 2024:266, [gpt-3/dataset_statistics/languages_by_word_count.csv](https://openai.com/research/gpt-3-dataset-statistics)). Despite its seemingly modest proportion, this still represents a vast volume of data, challenging the scale of traditional NLP-labeled corpora.

The contributions of the paper include:

- Methodology for extraction of lexical relationships using large language models
- Edge Dataset of extracted and partially evaluated lexical synonymy¹ relations
- Methodology for construction of lexical networks using graph models
- The Dash Cytoscape graph application (available at the development link: <http://liks.ffri.hr:8002>).

2. Previous Research and Theoretical Background

There have been numerous efforts to organize Japanese vocabulary on semantic principles since the first edition of thesaurus Word List by Semantic Principles (WLSP) (NINJAL 1964) and its revised edition (NINJAL 2004). It classifies Japanese lexicon into hierarchical semantic categories, enhancing understanding and teaching. Most recently, efforts have been made to develop more sophisticated corpora and searching tools e.g., BCCWJ (Maekawa et al., 2014).

¹ Dataset with extracted synonyms is available at: <https://docs.google.com/spreadsheets/d/13YdAdyecikMYV-pA6eQ8ZNibAgYRIZD5b/edit?gid=54417379#gid=54417379>

In addition to these efforts, there have been developments of the Japanese WordNet (Isahara et al., 2008; Goodman & Bond, 2021), which aligns with the structure of the English WordNet to illustrate semantic relationships such as synonymy and antonymy within the Japanese language. Many online resources, e.g., online dictionary *Goo jisho*, use Japanese WordNet. The EDR Electronic Dictionary also organizes Japanese vocabulary with semantic categorizations besides phonetic annotations (Takebayashi, 1993).

2.1 Applications of Graph Theory in Linguistics

The integration of semantic principles with computational methods naturally leads to the advancement of lexical graphs, which use graph theory to model complex relationships between words. The study of lexical graphs involves visualizing the relationships between lexical items, providing a structured and quantitative approach to understanding the lexicon (Vitevitch, 2008; Veremyev et al., 2019; Diessel, 2023). Lexical graphs are formed from nodes, representing words or lexemes, and edges, symbolizing the relationships or associations between these nodes (Perak & Kirigin, 2023). This method offers a visual and analytical framework to explore how words are interconnected within a language, helping uncover insights into the organization of the mental lexicon. For instance, graph analyses often reveal that lexical networks exhibit small-world characteristics, where high clustering coefficients and short path lengths enhance the efficiency of lexical retrieval and communication (Agustín-Llach & Rubio, 2024).

The construction of lexical graphs typically begins with the collection of lexical data. Classical natural language processing (NLP) techniques involve phases such as tokenization, lemmatization, and syntactic dependency tagging. Using these tools, lexical graphs can be constructed with edges denoting relationships determined by syntactic dependencies or collocations. For instance, the method presented by *Construction Grammar Conceptual Network* research group utilizes syntactic dependencies to create multilayer networks, which capture different lexical semantic features (Perak & Kirigin, 2023). These graphs facilitate tasks such as word sense disambiguation, sentiment analysis, and semantic search by providing a rich contextual framework for interpreting lexical data (Kirigin et al., 2021; 2022).

2.2 Facilitating Lexical Extraction With Large Language Models

Traditional NLP processing steps are challenging because they necessitate high computational resources and linguistic expertise to manage the subtleties of a linguistic task. Moreover, mapping syntactic features to semantic features introduces further complexity as it involves understanding broader context-dependent meanings, often not readily available from classical corpus methods.

Top tier Large language models (LLMs), like GPT-4 or GPT-4o significantly streamline the extraction and generation of lexical relationships by leveraging their extensive pre-training on diverse linguistic data. Key advantages can be summarized as follows:

- **Dynamic Generation:** LLMs can quickly produce multiple lexical relationships for any given word, reducing the need for manual data extraction and ensuring comprehensive coverage across diverse contexts (Zhang et al., 2023).
- **Contextual Relevance:** The model's ability to reason about lexical context (Perak et al., 2024) and generate contextually appropriate linguistic patterns reinforces relevant selection of the synonyms and antonyms.
- **Scalability:** Utilizing LLMs scales well with large numbers of queries, enabling the efficient processing of vast data inputs and the generation of detailed lexical networks without the exhaustive preprocessing required in traditional methods.

This modern generative approach allows for the extraction of rich lexical relationships, ensuring both comprehensive and contextually relevant lexical networks construction.

2.3 Graph Methods vs. Large Language Models: Rationale for Combining Approaches

While LLMs can generate extensive linguistic data, there are compelling reasons to combine graph methods. Graph methods provide a transparent and interpretable representation of lexical relationships, unlike the often-opaque nature of LLMs, where the internal representations are not easily accessible.

Secondly, graphs enable the visualization of complex lexical networks in a manner that is easy to comprehend. Using tools like Python's NetworkX or Igraph library, lexicographers can create visual representations that highlight the density and centrality of connections among lexemes. This visual format is particularly useful for identifying key semantic clusters, understanding the spread and influence of particular lexemes, and detecting patterns that might be obscured in the raw outputs of LLMs.

By employing graph methods, researchers and lexicographers can more effectively validate and fine-tune the lexical relationships generated by LLMs. Once the LLMs generate initial lexical relationships, these can be mapped onto a graph and examined for accuracy and relevance.

Furthermore, graphs allow for the incorporation of rich attributes for both nodes and edges. For instance, each node (lexeme) can store various attributes such as part of speech (POS), frequency, pre-2010 JLPT (Japanese Language Proficiency Test) level, and semantic domains, different attributes from diverse lexical resources, as well as different centrality measures. Edges (relationships) can include attributes like relational strength, type (synonym, antonym), and domain-specific context. This rich attribute management facilitates more nuanced analyses and the creation of robust lexical resources. Furthermore, graph methods excel at enabling easy updates and modifications. As new lexical data are generated or discovered, they can be seamlessly integrated into the existing graph structure.

In practical applications such as language learning tools, search engines, and natural language processing (NLP) applications, having a well-organized and transparent lexical network is invaluable. For example, language learners may benefit from visual aids that show the relationships between words, aiding their comprehension and retention (Bollegala et al., 2015; Tokuhiko, 2016). Similarly, NLP applications can leverage the structured data in graphs to improve algorithms for tasks like semantic search, text summarization, and sentiment analysis.

3. Methodology

The procedure for extracting lexical relations is illustrated in Figure 1 is detailed in the following section.

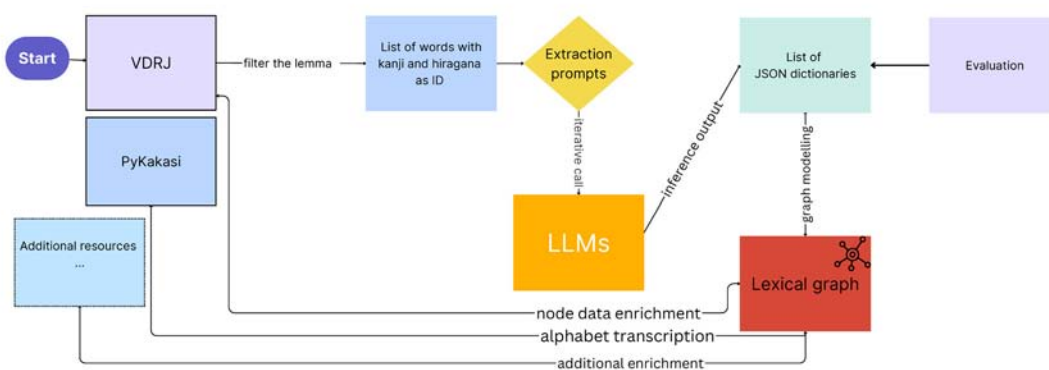


Fig. 1: Schematic representation of the lexical graph construction.

3.1 Lexicon Definition and Filtering

As a starting point for creating Japanese lexical network, we used *Vocabulary Database for Reading Japanese* (VDRJ) of over 60,000 Japanese words (Matsushita, 2011)², initially filtered to a subset of common nouns at the levels from N4 (the lowest) to N1 (the highest), according to pre-2010 JLPT scale³.

² See also Matsushita (2016) for building corpus-based vocabulary syllabus for Japanese language learners.

³ Pre-2010 JLPT scale corresponds to the current one in such a way that there is another level between levels 2 and 3 (see: https://jlpt.jp/e/reference/pdf/guide2011_e_02.pdf). Although the official correspondence between the JLPT scale and CEFR (Common European Framework of Reference for Languages: Learning, teaching, assessment) will not be available until 2025 (<https://www.jlpt.jp/e/cefrlevel/index.html>), some sources (Center for Language Education and Research Sophia University), provide the following correspondence of the levels of Japanese language courses to CEFR. (<https://www.sophia-cler.jp/study/ja/pdf/LevelsofJapaneseLanguageCourses.pdf>) JLPT5: A1-A2; JLPT4: A2; JLPT3: B1; JLPT2: B2; JLPT1: B2-C1.

The word list includes:

1. 3215 General common nouns (名詞 普通名詞一般), of pre-2010 JLPT levels N4 to N1.
2. 6200 general common nouns of N0 level (名詞-普通名詞-一般),
3. 1609 general common nouns that may function as verbal nouns (名詞-普通名詞-サ変可能),
4. 18 general common nouns that may function as nominal adjectives (-na adjectives) (名詞-普通名詞-形状詞可能),
5. 13 general common nouns that may function as adverbs (名詞-普通名詞-副詞可能),

Overall, we have included: 11055 unique lemmas for GPT processing.

3.2 Interfacing With GPT-4

The core of our methodology revolves around systematically interfacing the filtered lexicon with OpenAI's GPT-4o to extract lexical relationships. For this, we created a Python script iterating calls to the OpenAI API interface with a series of system prompts and precise query instructions to refine and structure the model's outputs. Each lexeme from the filtered list serves as a part of the input to GPT-4o, which is prompted to generate at least 15 synonyms and 5 antonyms for each item, as well as to infer their semantic domain.

For example, the lexeme '子供' (kodomo - child) would be introduced to GPT-4 with parts of prompts formulated as follows:

Here is a batch of lexemes (lemma) with their corresponding Part of Speech (POS) and katakana readings in Japanese.

The tasks are to: 1) Translate lexemes to English and 2) to propose lexical relations in Japanese.

For each lexeme in Japanese create a JSON dictionary with the following key: 'lexeme_data'. Inside this key determine: source_lemma, source_lemma_reading_hiragana, source_POS (from a list of {source_language_POS_list}).

Write data about translated lexeme with key: 'lexeme_translation'. Inside this key determine: target_language, target_lemma, target_POS.

Write the data about the synonym relation to other lexemes in Japanese with key: 'lexeme_synonyms'. Inside this key create a list of JSON dictionaries.

For each lemma determine at least 15 synonyms in Japanese, with their corresponding POS from a {source_language_POS_list}:

For each synonym determine strength of synonym relation as a float value in a range from 0.00 to 1.00.

Determine mutual sense and synonym domain in Japanese language and write this in Japanese.

Write data about the antonym relation to other lexemes in Japanese with the key: 'lexeme_antonyms'. Inside this key create list of dictionaries:

For each lempos determine at least 5 antonyms in {source_language}, with their corresponding POS from a {source_language_POS_list}:

For each antonym determine strength of antonymy relation as a float value in a range from 0.00 to 1.00. Where 0 refers to no synonymy and 1 refers to same meaning

Explain the antonymy relation and determine antonymy domain in {source_language} language and write this in {source_language}.

The prompt begins by introducing a batch of lexemes along with their corresponding Part of Speech (POS) and kana readings. This allows the GPT-4o to handle multiple lexemes in a structured manner. The tasks specified are two-fold: translating the lexemes to English as the target language and proposing lexical relations (synonyms and antonyms) in the source language.

The prompt's specificity and precision in query instructions are critical for minimizing ambiguity in the responses. By clearly defining the expected output format and the keys to be used in JSON dictionaries, the prompt guides GPT-4o to produce structured and uniform data⁴. The requirement to output data strictly in JSON format and avoid additional keys or characters ensures that the generated data is clean, easily parsable, and ready for subsequent analysis.

Furthermore, the output is rendered more controllable by setting the temperature parameter to 0. This parameter in the context of language models, such as GPT-4o, dictates the randomness of the generated text. A temperature of 0 ensures deterministic outputs, meaning the model always chooses the most probable next word, thereby reducing variability and enhancing consistency. This controllable output is particularly valuable in linguistic applications where precision and reliability are paramount, such as in the generation of synonyms and antonyms for lexical networks. By minimizing stochastic elements, the refined setting facilitates the creation of detailed and contextually accurate lexical relationships.

This structured approach supports efficient data manipulation and integration into lexical networks. Here's a simplified example demonstrating the output of the prompt:

⁴ With the introduction of the model GPT4o-2024-08-06 it is now possible to define strict structured output (see: <https://platform.openai.com/docs/guides/structured-outputs>).

Example JSON output for the input Lexeme: 子供 (こども):

```
{ "source_lexeme": { "lemma": "子供", "hiragana": "こども", "POS": "Noun" },
  "lexeme_translation": { "target_language": "English", "target_lemma": "child", "target_POS": "Noun" },
  "lexeme_synonyms": [ { "lemma": "児童", "hiragana": "こども", 'relation': 'synonymy', 'strength': 0.95, 'explanation': 'Both refer to young individuals.', 'synonymy_domain': '年齢', 'synonymy_domain_hiragana': 'じどう', 'synonymy_domain_translit': 'nenrei', 'synonymy_domain_translation': 'age', 'mutual_sense': '若い人', 'mutual_sense_hiragana': 'わかいひと', 'mutual_sense_translit': 'wakaihito', 'mutual_sense_translation': 'young person'}, ...],
  "lexeme_antonyms": [ { "antonym_lemma": "大人", "hiragana": "おとな", "POS": "Noun", "antonym_translation": "adult", "antonym_strength": 1.0, "antonymy_domain": "年齢", "antonymy_domain_hiragana": "ねんれい", "antonymy_domain_translation": "age", "antonymy_domain_explanation": "child and adult are opposites in terms of age", "antonymy_domain_translation": "age"}, ... ]
}
```

This example illustrates how the specified prompt leads GPT-4o to generate lexical data structured in JSON format.

3.3 Constructing Graph From JSON Output

To construct a graph object from the JSON output provided, we used Pandas and NetworkX Python libraries. The nodes in the graph represent lexemes (words), and the edges represent the relationships (synonyms and antonyms) between them with nodes and edges having specific attributes derived from the JSON data.

First, we parsed the JSON data to extract the relevant information about the source and target lexemes, i.e., synonyms, and antonyms. Importantly, the unique identifier for each node has to utilize both its kanji form and hiragana reading for a few critical reasons. (1) An orthographic form involving kanji may have multiple different words attached to it, e.g., 角 *kaku* 'square, angle', 角 *kado* 'corner', 角 *tsuno* 'horn' (anatomy). (2) Conversely, a word may have multiple orthographic variants - *tamago* 'egg' can be written using different kanji 卵 or 玉子, or can be written just in kana, i.e. たまご or タマゴ. (3) A word may have different nuances when written with different kanji, e.g., *furusato* 'one's hometown' is written 故郷 or 故里 in different contexts. (4) 2 words with different phonological forms may share both orthographic form and the meaning, comprising a pair of synonyms, e.g., 故郷 *furusato* and 故郷 *kokyoo* 'one's hometown'. As will be shown below, intricacies of the writing system pose problems when determining synonymy.

Subsequently, we constructed edges datasets of the lexical network graph from synonym or antonym relationships along with multiple attributes derived from the JSON data, including *synonym_strength*, *mutual_sense*, *synonymy_domain*, *synonymy_explanation*.

The *synonym_strength* attribute quantitatively indicates the degree of semantic similarity between the source lexeme and its synonym, with a higher value indicating a closer semantic match. The *mutual_sense* with its hiragana transcript and English translation offers a shared contextual meaning, providing additional semantic context for the relation. Moreover, *synonymy_domain* and its hiragana and translation variants classify the categorical nature of the lexical relationship ensuring domain-specific clarity. The *synonymy_explanation* gives a qualitative articulation of why the terms are considered synonyms.

Edges representing antonyms are constructed with similar attributes to capture the oppositional nature of the lexical relationships. Attributes such as *antonym_strength*, *antonymy_domain*, *antonymy_explanation* are essential in this context.

The *antonym_strength* attribute provides a quantitative measure of the strength of the antonym relationship, with a value of 1.0 indicating complete oppositeness. The *antonymy_domain* attribute categorizes the nature of this opposition by contextual domain, such as 年齢 *nenrei* ‘age,’ which in this case is further clarified with its hiragana representation. The *antonymy_explanation* offers a qualitative description, for instance, detailing that “child and adult are opposites in terms of age”.

3.4 Graph Enrichment

The initial lexical network graph, constructed using GPT-4o outputs, represents a complex relational structure of Japanese lexemes. To further enrich this graph with comprehensive lexical information, the original Vocabulary database for reading Japanese (Matsushita 2011) has been utilized to incorporate additional linguistic attributes:

- Standardized Frequency per million in 10 Written Domains (Fw) that provides a quantitative measure of how commonly each lexeme is used.
- Pre-2010 JLPT Level: The proficiency level of the lexeme.

The Jisho API (<https://jisho.org/>) has been employed providing a new JLPT scale where available as additional lexical attributes.

To facilitate the understanding and learning of Japanese for non-native speakers we inserted alphabet (*romaji*) transcriptions from hiragana using PyKakasi, a comprehensive Python library available on GitHub (<https://github.com/miurahr/pykakasi>).

3.5 Graph Analysis After Creating the Lexical Network

The network is composed of 137,870 edges representing synonymy relationships between lexemes and 54,324 edges representing antonymy relationships. In this lexical network 60,750 nodes are derived from synonym edges. When additional semantic dimensions such as mutual sense and domain are considered, the number of nodes connected by synonym edges expands to 99,068.

In terms of antonymy, the network includes 19,412 nodes connected by antonym edges. When domain nodes are included, the count of nodes linked by antonymy increases to 20,360.

The network thus contains a total of 65,807 nodes that are connected by either synonymy or antonymy edges, and inclusive with mutual sense and domain, the number of nodes rises to 104,427.

Table 1: Nodes and edges statistics for extracted Lexical Network

Type	Count
Nodes from synonym edges	60750
Nodes from synonym edges including mutual_sense and domain	99068
Nodes from antonym edges	19412
Nodes from antonym edges including mutual_sense and domain	20360
Nodes from synonym and antonym edges	65807
Nodes from synonym antonym edges including mutual_sense and domain	104427
Synonym edges	137870
Antonym edges	54324

4. Visual Representation Using Dash Cytoscape

To effectively visualize the complex relationships within Japanese lexical networks, Dash Cytoscape (<https://dash.plotly.com/cytoscape>) a tool for interactive graph visualizations, has been utilized. Visualization involves generating elements that represent both nodes and edges with detailed attributes (see Fig 1.). The app is currently running on a development server with address <http://liks.ffri.hr:8002> .

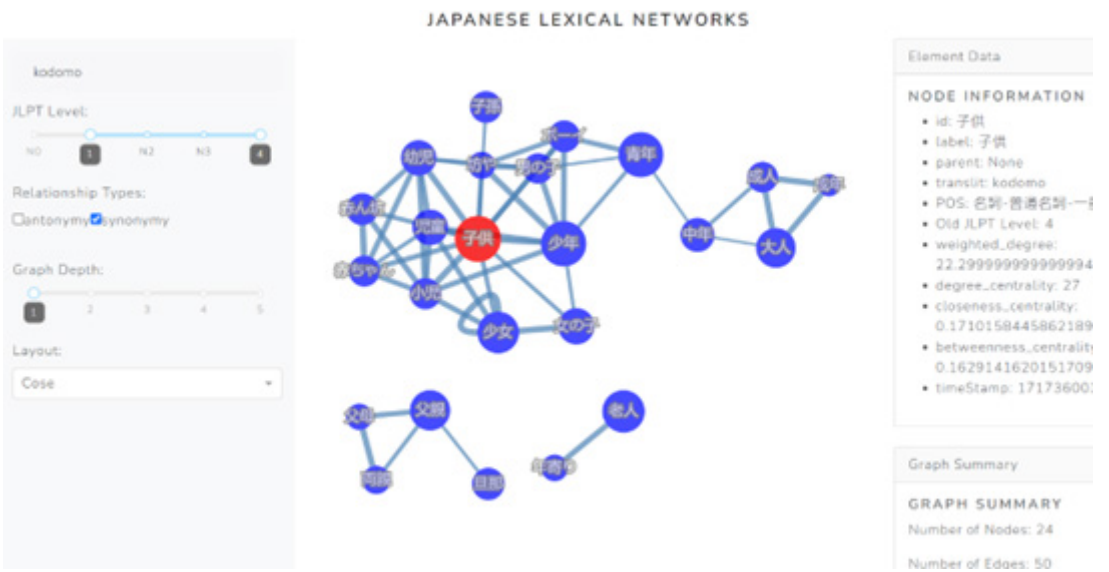


Fig. 2: Japanese Lexical networks example illustration for the lexeme ‘子供’ *kodomo* ‘child’ filtered for lexemes pre-2010 JLPT level (1-4) and synonym relations in graph depth 1 using layout Cose.

The interactive features provided by Dash Cytoscape facilitate user engagement and deeper exploration of the lexical network:

- Search Functionality: Users can search for specific lexemes by their transliteration or node identifiers, including kanji or hiragana.
- Filter and Range Selection:
 - JLPT levels: This slider allows users to filter lexemes based on their pre-2010 JLPT level. Additionally, there is also ‘N0 level’, i.e., items not included in JLPT.
 - Relationship Type Selector: Users can choose between Synonymy, Antonymy and Domain types of lexical relationships.
 - Graph Depth: This slider lets users control the depth of the network graph. Increasing depth shows more layers of connected lexemes.
- Dynamic Layouts: Multiple layout options (e.g., breadthfirst, grid, circle, cose) can highlight various properties of the graph structure, such as centrality or clustering.

Filters enable users to customize which nodes and relationships are displayed based on their proficiency level and interest.

5. Evaluation

In this chapter, output items will be evaluated and some shortcomings of the model will be pointed out. The scalar approach to defining synonymy relationships—where similarity between word senses is quantified by a *synonym_strength* attribute ranging from 0 to 1—inevitably encompasses a broad spectrum of semantically related entities. However, during the evaluation, we recognized that this broad categorization could lead to oversimplifications, failing to capture the nuanced distinctions between different types of synonymy. To address this, we introduced additional Relation Type within synonym pairs.

Let us first consider key concepts for evaluating synonymy. Synonymy is usually defined as a phenomenon whereby a single meaning is associated with more than one distinct lexical item. There have been numerous attempts to identify near synonyms, e.g., Edmonds & Hirst (2002), Inkpen & Hirst (2003), Terada & Yoshida (2007). Most recently, Ishii & Sasaki (2024) proposed a method of synonym identification by mapping word meanings using dictionary definitions and Sentence-BERT. Arguably, fully synonymous words do not exist or are very rare (Lyons, 1968, p. 448). Thus, it is a common practice to combine synonyms and near synonyms in the same category, see *Word List by Semantic Principles* (NINJAL 2004, p. 13). Partial synonymy is a type of synonymy relating to a certain meaning, e.g., Japanese words *kotoba* ‘language; word’ and *go* ‘word’ are partly synonymous. Examples are also *hot* and *spicy* in English, (Taylor, 2002, p. 266), and *joozu* ‘skillful’ and *umai* ‘skillful’ in Japanese (Ishii & Sasaki,

2004, p. 2993). Near synonyms (NS), also called plesionyms by Cruse (1986, p. 285), are words similar in meaning, which tend not to be contrastive, but are used in different contexts (Taylor, 2002, p. 263). For convenience, Partial and Near Synonyms will be included in the same category, **Near Synonyms (NS)**.

Topic-related items are those which occur within the same or closely related semantic domain. Subtypes of TR items are following:

1. **TR_hyponyms** e.g., *sandaru* ‘sandals’ is a hyponym for *kutsu* ‘shoes’,
2. **TR_hyponyms** e.g., *idoo* ‘movement’ is a hypernym for *ryokoo* ‘travel’,
3. **TR_meronyms** e.g., *shokutaku* ‘dining table’ is a meronym for *daidokoro* ‘kitchen’ (Winston 1987: 421)
4. **TR_holonyms**, e.g., *kyoouiku* ‘education’ is holonym for *benkyoo* ‘studying’,
5. **TR_class** e.g., *kobachi* ‘small bowl’ belongs to the same class as *sara* ‘plate’,
6. **TR_object with an attribute** e.g., *akajiso* ‘red perilla (botany)’ has an attribute *aka* ‘red’,
7. **TR_possession** e.g., *tokeishokunin* ‘clockmaker’ for *tokei* ‘clock’,
8. **TR_subject of activity**, e.g., *fuufu* ‘married couple’ for *kekkon* ‘marriage’
9. **TR_topology**, e.g., *obentooya* ‘bento shop’ for *bentoo* ‘bento (lunchbox)’,
10. **TR_paraphrase**, e.g., *ushi no chichi* ‘cow’s milk’ for *gyuunyuu* ‘(cow’s) milk’.

Other labels are following:

ERROR_kanji_based is a type of relation where a target word is apparently listed only because its first kanji is the same as the source word but they are semantically unrelated, e.g., 角, with specified hiragana reading *kaku* ‘corner, angle’ gave words such as 角笛 *tsunobue* ‘horn (music instrument)’ and 角膜 *kakumaku* ‘cornea (anatomy)’. This is probably related to the fact that the same grapheme 角 has 3 words attached to it *kaku* ‘angle, square’, *kado* ‘corner’ and *tsuno* ‘horn (anatomy)’.

ERROR_verb was an instance of a wrong POS, namely, verb *kaimotomeru* ‘to buy’ was listed as a synonym for verbal noun *kaimono* ‘shopping’. **ERROR_Chinese**: in one case, as a synonym for 言葉 *kotoba* ‘word, language’, “言” was listed which does not exist in Japanese. It could be the Chinese homograph meaning ‘word’. **ERROR_same** are instances where the source word was just repeated as the target.

Cases of multiple orthographic forms are listed as **Orth_var**. Such problems have been discussed in corpus linguistics (see Kuroda et al. 2011). If an item represents an orthographic variation, it should not be regarded as a synonym. E.g., source word 御腹 *onaka* ‘stomach’ has also orthographic forms お腹, which the model lists

as synonym of a strength 0.9, lower than 腹 *hara* ‘belly’ with 0.95. This problem should be addressed in the subsequent models.

Let us now turn to more specific evaluation details. The evaluation dataset comprises a subset of synonym edges extracted from a larger lexical network focused on Japanese vocabulary related to the pre-2010 JLPT Level 4. This sample includes 1085 edges with 84 source lemmas that were randomly selected. These lemmas represent a diverse range of basic Japanese vocabulary, from everyday nouns like 人 *hito* ‘person’ to more specific terms like 結婚 *kekkon* ‘marriage’. Each source lemma has an average of 12.5 synonyms that altogether create a set with 1050 unique synonym lemmas.

The evaluation process, conducted manually by one of the authors, involved a systematic review of each word pair and generated metadata, using resources such as online dictionary *Goo jisho* <https://dictionary.goo.ne.jp/> which includes Shogakukan Thesaurus (2003) and NINJAL (2004). The evaluator initially assessed each semantic relation with respect to the enlisted domain using numeric scores on a scale from 1 to 5, with 1 indicating poor and 5 excellent quality. The scores were then analyzed to produce statistical measures for each type of relationship within the dataset.

- EVALUATION_synonym: represents the assessment of the quality of the synonym relationship. The results show a mean score of 4.08 with a standard deviation of 0.70. The majority of evaluations are high, with 75% of scores being 4 or higher.
- EVALUATION_sense evaluates the alignment of mutual senses between the synonym pairs. It shows a high level of agreement, with a mean score of 4.97 and 75% of scores being 5.
- EVALUATION_domain assesses the domain consistency of the synonym pairs, with a mean score of 4.91 and a standard deviation of 0.49. Similar to the sense evaluation, the scores indicate strong domain alignment.

The synonym relationships in the evaluation set are categorized according to their types described above. The distribution of these relation types is represented in the Table 2:

Table 2: Relation types per category. Relation Type is a proposed class of semantic relationship between synonym pairs. Count is the number of synonym pairs analyzed for each relation type. Mean Strength represents the average strength of synonym relationships within each relation type. Std Dev is the standard deviation of synonym strength, indicating the variability or spread of the synonym strength values within each relation type. Min Strength is the minimum synonym strength observed within each relation type. Max Strength is the maximum synonym strength observed within each relation type. Correlation (Synonym Strength, Evaluation Synonym) represents the correlation between synonym strength and the evaluation score.

Relation Type	Count	%	Mean Strength	Std Dev	Min Strength	Max Strength	Correlation
TR_MERONYM	500	46.1	0.542080	0.18603	0.150000	0.920000	0.090661
NS	215	19.8	0.836419	0.10769	0.350000	0.980000	0.222864
TR_HYPONYM	133	12.3	0.587218	0.17512	0.100000	0.950000	0.112714
TR_CLASS	132	12.2	0.564773	0.18520	0.150000	0.880000	0.106177
TR_HYPERNYM	32	2.9	0.657812	0.230045	0.200000	0.950000	0.177710
ERROR, KANJI BASED	16	1.4	0.490000	0.172378	0.200000	0.800000	NaN
TR_OBJECT_WITH_ATTRIBUTE	13	1.2	0.419231	0.137747	0.200000	0.650000	-0.532802
TR_HOLONYM	11	1.0	0.613636	0.130558	0.400000	0.800000	0.034641
ERROR, SAME	8	0.7	0.912500	0.210017	0.400000	1.000000	NaN
TR_SUBJECT_OF_ACTIVITY	5	0.5	0.680000	0.115109	0.500000	0.800000	NaN
ORTH_VAR	4	0.4	0.925000	0.064550	0.850000	1.000000	NaN
ERROR, SEMANTIC	4	0.4	0.600000	0.070711	0.550000	0.700000	-0.471405
TR_POSSESSION	4	0.4	0.537500	0.262599	0.150000	0.700000	-0.714545
TR_TOPOLOGY	3	0.3	0.516667	0.275379	0.200000	0.700000	-0.419314
TR_PARAPHRASE	2	0.2	0.650000	0.070711	0.600000	0.700000	NaN
ERROR, VERB	1	0.1	0.650000	NaN	0.650000	0.650000	NaN
ERROR, CHINESE	1	0.1	0.750000	NaN	0.750000	0.750000	NaN
ERROR_STRUCTURAL	1	0.1	0.600000	NaN	0.600000	0.600000	NaN

Near Synonyms comprise 19.8% and Topic related words, including all 10 subtypes, account for 77.1% in the dataset. NS has a highest mean strength (0.836), indicating that GPT self evaluation coincides with classification of the synonym pairs as generally closely related, with a moderate correlation (0.222864) to the evaluation score. TR (Topic-Related) items show lower mean strengths (from 0.542 to 0.419) and lower correlations to the evaluation score, suggesting these relationships are more diverse in nature.

The model correctly lists words with strong synonymy relation. For instance, for source word *ryoori* ‘cooking, cuisine, dish’ GPT lists target words *choori* ‘cooking’, *suiji* ‘cooking’, and *kukkingu* ‘cooking’; and also *shokuji* ‘meal’ and *tabemono* ‘food, dish’, partial synonyms on the meaning ‘dish’ which we all classified as NS.

Let us consider examples of mutual sense and synonymy domain. Item *otokonoko* ‘boy’, has synonymy_strength 0.95 with *shoonen* ‘boy’ and is labeled near synonym;

their mutual_sense is *wakai dansei* ‘young male’ and synonymy_domain *nenrei* ‘age’; next item is *kodomo* ‘child’, strength 0.85, whose mutual_sense is *wakai hito* ‘young person’ and synonymy_domain also *nenrei* ‘age’. Next 2 items, *danshi* ‘male child’ (0.88) and *danji* ‘male child’ (0.75) have both mutual_sense *dansei no kodomo* ‘male child’ with synonymy_domain *seibetsu* ‘gender’; *jidoo* ‘pupil’ (0.8) has mutual_sense *gakkoo ni kayou kodomo* ‘school-going child’ with synonymy_domain *kyoouiku* ‘education’ and synonymy_explanation “Both refer to children, typically of school age.”

The model correctly differentiates cases when a lexeme has 2 POS’s, e.g., noun and nominal adjective: *genki* ‘vigorously, healthy’ and ‘vigor, health’ yield 2 types of synonyms, *kappatsu* ‘lively’ *kappatsusa* ‘liveliness’.

There are cases when GPT does not list multiple meanings of a word. E.g., *midori* has 2 main meanings – ‘green color’, ‘greenery’. However, the model captures only the second one giving topic related words e.g., *kusa* ‘grass’, *ha* ‘leaves’, *mori* ‘wood’. This is not the case with e.g., *aka* ‘red color’ where various nuances of red are enumerated (however, it has no multiple meanings). This could be mitigated by using a more sophisticated prompt.

As for the number of NS for each entry, there is room for improvement. For example, 菓子 *kashi* ‘confectionary’ includes a synonym *okashi* which differs from source only by polite prefix *o-*, 2 near synonyms *suiitsu* ‘sweets’ and *dezaato* ‘dessert’ and the next 11 words are hyponyms e.g., *kyandii* ‘candy’, *chokoreeto* ‘chocolate’ etc. Words such as *kanshoku* ‘food eaten between meals’ and *oyatsu* ‘snack’ are conspicuously absent although we would expect to find them here. Also, it is noticed that all hyponyms are items with sweet taste despite the fact that *okashi* includes savory snacks, too. In the future research, a more systematic comparison with available resources e.g., Japanese WordNet will be considered.

It should be added that lists of near synonyms and related words may be relevant for Japanese language education, e.g., developing vocabulary syllabus, since a teaching unit usually involves specific topic(s). Yamauchi (2013) gives a comprehensive overview on the relation between a topic and the choice of vocabulary, and research on the recently developed *Japanese Topic-Oriented Conversation Corpus (J-TOCC)* (Nakamata et al. 2023) shows such relation even more specifically.

6. Discussion

6.1 Synergy of AI Tools and Traditional Lexicography

The integration of AI tools, specifically large language models such as GPT-4o, with traditional lexicographic methods marks a significant advancement in the field of lexicography (De Schryver, 2023; Lew, 2024).

AI tools, such as GPT-4o, revolutionize this process by automating the extraction and generation of lexical relationships. The collaborative approach leverages the computational strengths of AI with the nuanced understanding of human expertise.

It extends the applicability of lexicographic resources across languages and dialects, making it possible to construct comprehensive lexical databases or for less-studied languages and dialects more quickly than traditional methods alone would allow. Additionally, this synergy supports the development of advanced language learning tools and natural language processing applications, which require robust and detailed lexical databases to function effectively.

In addition, LLMs like GPT-4o can be viewed as an extension of linguistic corpora. These models are pre-trained on vast and diverse datasets that encapsulate a wide array of linguistic patterns, usages, and contexts. However, it is crucial to acknowledge that the output generated by these models, especially in lexical items, does not always equate to words (i.e. long word units) as defined in standardized linguistic resources such as UniDic used in e.g., the Balanced Corpus of Contemporary Written Japanese (BCCWJ) (Maekawa et al., 2014). In contrast, LLMs might generate lexical items based on statistical probabilities and patterns observed during training, which may include neologisms, colloquialisms, or less standard forms.

6.2 Refining LLM Prompts With Relationship Types

A potential avenue for further research involves refining LLM prompts to explicitly incorporate the identified relationship types. By providing specific prompts that include terms such as “near synonyms,” “hyponyms,” or “meronyms,” we can potentially elicit more precise and nuanced lexical relationships from the model. For instance, instead of a general prompt requesting synonyms, a more specific prompt could ask for “near synonyms with a similar level of formality” or “hyponyms of the target word.” This level of granularity could significantly enhance the quality and specificity of the extracted lexical data, leading to more accurate and informative lexical networks.

Additionally, exploring the potential of combining multiple LLMs or fine-tuning existing models on specialized lexical datasets could further improve the accuracy and depth of the generated lexical relationships. This could involve training LLMs on curated corpora of lexical data, including manually annotated examples of various relationship types.

6.3 Adaptability to Other Languages or Different LLMs

The methodology presented is adaptable to other languages and LLMs. By applying the same core principles and leveraging multilingual LLMs, detailed lexical networks can be constructed for various languages. This involves tailoring the lexicon to the target language and utilizing a suitable LLM to generate synonyms, antonyms, and other lexical relationships. The flexibility to employ different LLMs enhances the potential for creating nuanced and culturally specific lexical networks (Huang et al., 2023).

7. Conclusion

This study proposes a novel method for constructing Japanese lexical networks by LLMs such as GPT-4o. The approach involves:

1. **Extracting Lexical Relationships:** GPT-4o is used to generate synonyms, antonyms, and domain information for Japanese words.
2. **Constructing Lexical Networks:** A graph is built with nodes representing words and edges representing relationships (synonyms/antonyms) with attributes like strength and domain.
3. **Enriching the Network:** Additional information like frequency and JLPT level is added from external resources.
4. **Interactive Visualization:** Dash Cytoscape allows users to explore the network by searching, filtering, and customizing layouts.

Evaluation of a sample set showed high accuracy for synonym relationships (mean score 4.08) and strong alignment of sense and domain. The study acknowledges limitations of a single “synonym strength” measure and proposes further classification of synonym types (e.g., near synonyms, topic-related). It was established that in the sample dataset Near Synonyms comprise 19.8% and Topic related items, including all 10 subtypes, account for 77.1%. Future work will focus on using these identified relation types to refine LLM prompts, aiming to improve the precision and accuracy of generated lexical relationships.

This approach offers a scalable and adaptable framework for building comprehensive lexical networks, with significant implications for computational linguistics and NLP applications.

References

- Agustín-Llach, M. P., & Rubio, J. (2024). Navigating the Mental Lexicon: Network Structures, Lexical Search and Lexical Retrieval. *Journal of Psycholinguistic Research*, 53(2), 21.
- Aotani, N., & Takahashi, S. Y. (2021). An analysis of Japanese EFL learners’ lexical network changes. In *Frontiers in Education* (Vol. 5, p. 621437). Frontiers Media SA.
- Ban Kirigin, T., Bujačić Babić, S., & Perak, B. (2021). Lexical sense labeling and sentiment potential analysis using corpus-based dependency graph. *Mathematics*, 9(12), 1449.
- Ban Kirigin, T., Bujačić Babić, S., & Perak, B. (2022). Graph-Based Taxonomic Semantic Class Labeling. *Future Internet*, 14(12), 383.
- Bollegala, D., Maehara, T., Yoshida, Y., & Kawarabayashi, K. I. (2015). Learning word representations from relational graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 29, No. 1).
- Cruse, D.A. (1986). *Lexical Semantics*. Cambridge University Press
- De Schryver, G. M. (2023). Generative AI and Lexicography: The Current State of the Art Using ChatGPT. *International Journal of Lexicography*, 36(4), 355–387.
- De Schryver, G. M. (2023). Generative AI and Lexicography: The Current State of the Art Using ChatGPT. *International Journal of Lexicography*, 36(4), 355–387.

- Diessel, H. (2023). *The construction: Taxonomies and networks*. Cambridge University Press.
- Edmonds, P., & Hirst, G. (2002). Near-Synonymy and Lexical Choice, *Computational Linguistics*, 28(2), 105–144.
- Goodman, M. W., & Bond, F. (2021). Intrinsically interlingual: The Wn Python library for wordnets. In: *Proceedings of the 11th Global Wordnet Conference* (pp. 100–107).
- Huang, S., Mei, Y., Wang, Y., Yang, Z., & Gong, S. (2023). *Ranking of Large Language Model (LLM) Cultural Bias*. World Association of Artificial Consciousness.
- Inkpen, D., & Hirst, G. (2003). Near-Synonym Choice in Natural Language Generation. In N. Nicolov, K. Bontcheva, G. Angelova, & R. Mitkov (Eds.), *Recent Advances in Natural Language Processing III: Selected papers from RANLP 2003*.
- Isahara, H., Bond, F., Uchimoto, K., Utiyama, M., & Kanzaki, K. (2008). Development of the Japanese WordNet. *Proceedings of the Sixth International Conference on Language Resources and Evaluation*. (pp. 2420–2423).
- Ishii, Y., & Sasaki, M. 2024. Sentence-BERT to gogi teigibun o riyoo shita gogikan no ruigi hantei shuhoo [Synonym Detection between Senses of Japanese Words Using Sentence-BERT and Sense Definitions Sentence-BERT]. *Shizen gengo shori gakkai happyoo ronbunshuu* (The Association for Natural Language Processing).
- Kovács, L., Bóta, A., Hajdu, L., & Krész, M. (2021). Networks in the mind—what communities reveal about the structure of the lexicon. *Open linguistics*, 7(1), 181–199.
- Kuroda, K., Kuribayashi, T., Bond, F., Kanzaki, K., & Isahara, H. (2011). Nihongo waadonetto no ihyooki taioo to heikoo koopasu e no gogi tagudzuke. [Orthographic variants and multilingual sense tagging with the Japanese WordNet]. In *17th Annual Meeting of the Association for Natural Language Processing*, p. A4–1, Toyohashi.
- Lew, R. (2024). Dictionaries and lexicography in the AI era. *Humanities and Social Sciences Communications*, 11(1), 1–8.
- Lyons, J. (1968). *Introduction to Theoretical Linguistics*. London: Cambridge University Press.
- Maekawa, K., Yamazaki, M., Ogiso, T., Maruyama, T., Ogura, H., Kashino, W., & Den, Y. (2014). Balanced corpus of contemporary written Japanese. *Language resources and evaluation*, 48, 345–371.
- Matsushita, T. (2011). *Nihongo o yomu tame no goi deetabeesu, Ver. 1.11* [The Vocabulary Database for Learners of Japanese, Version 1.11] Retrieved December 7, 2023, from <http://www17408ui.sakura.ne.jp/tatsum/database.html#vdrj>.
- Matsushita, T. (2016). Koopasu shutsugendo kara mita goi shirabasu [Vocabulary syllabus based on corpus frequency]. In A. Mori, & K. Shuppan (Eds.), *Niizu o fumaeta goi shirabasu* (pp. 53–77).
- Nakamata, N., Ota, Y., Kato, E., Sawada, H., Shimizu, T., & Mori, A. (2023). “Nihongo wadaibetsu kaiwa koopasu: J-TOCC” no kaisetsu [Introduction to Japanese Topic-Oriented Conversation Corpus: J-TOCC]. In Nakamata, N. (Ed.), *Wadaibetsu koopasu ga hiraku nihongo kyooiku to nihongogaku* (pp. 15–51). Hituzi syobo.

- Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)*, 41(2), 1–69.
- NINJAL (National Institute for Japanese Language and Linguistics). (2004). *Bunruigoihyoo – zooho kaiteiban* [Word List by Semantic Principles – Revised and Enlarged Edition], Dainippon tosho
- Oliveira, H. G., & Rodrigues, R. (2023). GPT3 as a Portuguese Lexical Knowledge Base?. In *Proceedings of the 4th Conference on Language, Data and Knowledge* (pp. 358–363).
- OpenAI. (2023). GPT-4 Technical Report. *ArXiv:2303.08774*
- Perak, B., & Kirigin, T. B. (2023). Construction Grammar Conceptual Network: Coordination-based graph method for semantic association analysis. *Natural Language Engineering*, 29(3), 584–614.
- Perak, B., Beliga, S., & Meštrović, A. (2024). Incorporating Dialect Understanding Into LLM Using RAG and Prompt Engineering Techniques for Causal Commonsense Reasoning. In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)* (pp. 220–229).
- Takebayashi, Y. (1993). EDR electronic dictionary. In *Proceedings of Machine Translation Summit IV* (pp. 117–126).
- Taylor, J. R. (2002). Near synonyms as co-extensive categories: ‘high’ and ‘tall’ revisited. *Language Sciences*, 25, 263–284.
- Terada A., & Yoshida M. (2007). A System for Constructing a Synonym Dictionary, *Journal of Natural Language Processing Journal of Natural Language Processing*, 15(2), 39–58.
- Tokuhiro, Y. (2016). Goimappu o mochiita shokyu kanji goi kyoozai no kaihatsu. [Development of beginner’s kanji vocabulary teaching materials using vocabulary maps]. *JSL Kanji gakushuu kenkyuukai-shi*, 8, 15–19.
- Veremyev, A., Semenov, A., Pasiliao, E. L., & Boginski, V. (2019). Graph-based exploration and clustering analysis of semantic spaces. *Applied Network Science*, 4, 1–26.
- Vitevitch, M. S. (2008). What can graph theory tell us about word learning and lexical retrieval?. *Journal of Speech, Language, and Hearing Research*, 51(2). [https://doi.org/10.1044/1092-4388\(2008/030\)](https://doi.org/10.1044/1092-4388(2008/030))
- Winston, M. E., Chaffin, R., & Herrmann, D. (1987). Taxonomy of Part-whole relations, *Cognitive Science*, 11, 417–444.
- Yamauchi, H. (Ed.) (2013). *Jissen nihongo kyooiku standaado* [Practical Japanese Language Education Standards], Tokyo: Hituzi shobo
- Zhang, Z., Li, B., & Zeng, Z. (2023). Review of research on synonym equivalence relation mining. In: *2023 International Conference on Intelligent Education and Intelligent Research (IEIR)* (pp. 1–6). IEEE.

Zhang, Z., Pu, P., Han, D., & Tang, M. (2018). Self-adaptive Louvain algorithm: Fast and stable community detection algorithm based on the principle of small probability event. *Physica A: Statistical Mechanics and Its Applications*, 506, 975–986.

Acknowledgements

This work was supported by projects: a) UNIRI CLASS1 A1 Mikrokvalifikacije: Jezične tehnologije i digitalna obrada teksta (JTDOT), b) UNIRI PROJEKTI ISKUSNIH ZNANSTVENIKA 2023: Matematičko modeliranje u obradi prirodnog jezika (uniri-iskusni-prirod-23-150). This paper was prepared as part of the institutional project *Japanese language and society: basic concepts and empirical analysis* (10/1/2023-9/30/2027), Juraj Dobrila University of Pula, Croatia.

Contact information

Dragana Špica

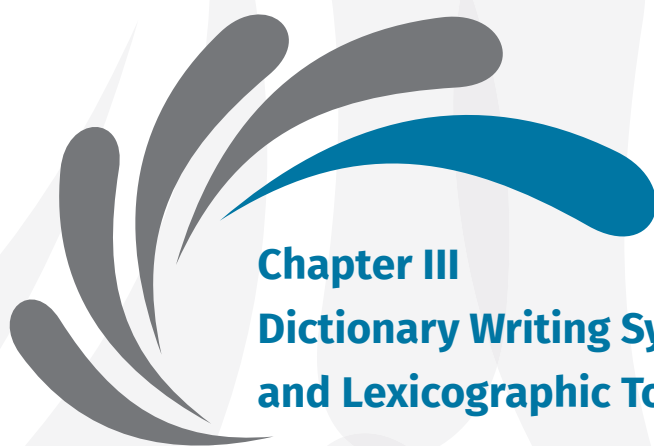
Sveučilište u Puli

dragana.spica@unipu.hr

Benedikt Perak

Sveučilište u Rijeci

bperak@uniri.hr



Chapter III
Dictionary Writing Systems
and Lexicographic Tools

Michaela Denisová, Gilles-Maurice de Schryver, and
Pavel Rychlý

THE AUTOMATIC DETERMINATION OF TRANSLATION EQUIVALENTS IN LEXICOGRAPHY: WHAT WORKS AND WHAT DOESN'T?

Abstract Cross-lingual embedding models act as facilitator of lexical knowledge transfer and offer many advantages, notably their applicability to low-resource and non-standard language pairs, making them a valuable tool for retrieving translation equivalents in lexicography. Despite their potential, these models have primarily been developed with a focus on Natural Language Processing (NLP), leading to significant issues, including flawed training and evaluation data, as well as inadequate evaluation metrics and procedures. In this paper, we introduce cross-lingual embedding models for lexicography, addressing the challenges and limitations inherent in the current NLP-focused research. We demonstrate the problematic aspects across three baseline cross-lingual embedding models and three language pairs and outline possible solutions. We show the importance of high-quality data, advocating that its role is vital compared to algorithmic optimisation in enhancing the effectiveness of these models.

Keywords translation equivalent determination; cross-lingual embedding models; evaluation

1. Introduction

In computational lexicography, translation equivalent retrieval is a non-trivial task with many challenges, including language ambiguity, variety, and disparity. Pinpointing translation equivalents relies heavily on human intuition and is time-consuming. An alternative approach is to use fully-automated Natural Language Processing (NLP) methods and algorithms, speeding up the process by offering a set of translation-equivalent candidates for a headword.

Cross-lingual embedding models represent a category of methods suitable for this task. These models are appealing and beneficial for lexicography for the following reasons. Firstly, as opposed to their parallel-data-based counterparts, they utilise *comparable data*, which are thus also available for under-resourced and non-standard language combinations. Secondly, the text variety of comparable data is usually much wider than in parallel corpora, offering more natural equivalents and vocabulary.

Despite these advantages, the research that has been done on these models is typically NLP-focused, improving the optimisation of the algorithms and technical aspects of these models while neglecting the lexicographic point of view (Ruder et al., 2019). The problems that arise from this, range from incorrect terminology usage to erroneous training and evaluation data and inappropriate evaluation metrics and methods, impeding our ability to interpret the results correctly and monitor the progress accurately.

In this paper, we introduce cross-lingual embedding models for translation equivalent retrieval in lexicography. We provide new insights about the training data, evaluation data, methods and metrics. We show how the selection of training data influences the results, analyse the widely used evaluation datasets and assess evaluation methods and metrics on three example language pairs: English-Slovak, English-Estonian, and English-Korean. We point out the problems and challenges and outline possible solutions. We show why the currently-used training data and evaluation data are faulty, why the evaluation method and metrics are inadequate, and how each of these can be improved.

Our motivation is to provide valuable material for lexicographers about cross-lingual embedding models for translation-equivalent retrieval that could serve as a stepping stone in their bilingual dictionary projects. Moreover, we emphasise the importance of high-quality data, which is often more crucial for achieving desired outcomes than the optimisation of the underlying algorithm.

The article is structured as follows. Section 2 outlines the background of cross-lingual embedding models. Section 3 compares two sources of the training data and demonstrates the results across three baseline models. Section 4 analyses evaluation datasets for three language pairs and shows how to allow for inaccuracies in them during evaluation. Section 5 discusses the most common evaluation metric and suggests an alternative. Section 6 offers concluding remarks.

2. Cross-Lingual Embedding Models

Word embeddings, or vector representations of words, are used to mathematically express a word's meaning based on the context words accompanying the word (Smith, 2019). In recent years, cross-lingual embedding models have appealed to many researchers mainly due to their ability to connect meanings across languages. These models aim to align two (or more) sets of separately trained monolingual word embeddings into a shared space where similar words obtain similar vectors (Ruder et al., 2019).

Except for monolingual word embeddings, they could involve a certain level of supervision in training. Based on the size of the training datasets in a word-to-word format, the models can be categorised into supervised (usually up to 5,000 headwords in the training dataset), semi-supervised (a small training dataset or mode that relies on identical strings and numerals occurring in monolingual word embedding vocabulary), and unsupervised (no training dataset) (Ruder et al., 2019).

In NLP, translation equivalent retrieval is referred to as the bilingual lexicon induction task (BLI) or the bilingual dictionary induction task (BDI). From a lexicographic point of view, this terminology is incorrect since “lexicon” or “dictionary” have much broader meanings and are not limited to translation equivalents only.

In the BLI task, the main objective is to find the most suitable translation equivalent(s) for a headword, typically by computing cosine similarity between the vectors of the

headword and the translation-equivalent candidates. Afterwards, the retrieved list of translation-equivalent candidates is compared to the evaluation dataset, a word-to-word list (Ruder et al., 2019).

Additionally, precision at k ($P@k$) is the most common reported metric, where k represents the number of translation-equivalent candidates retrieved for a headword. The precision computes the ratio of the correctly retrieved translation equivalents to all retrieved translation-equivalent candidates. Other crucial metrics, such as recall and F1 score, are less frequently employed. Recall measures the ratio of the correctly retrieved translation equivalents to the word pairs occurring in the evaluation dataset, while the F1 score captures both metrics, precision and recall, showing the balance between them.

In this paper, we selected the three most cited baseline cross-lingual embedding models for our experiments: MUSE (Conneau et al., 2018), VecMap (VM) (Artetxe et al., 2018a; b), and RCLS (Joulin et al., 2018). All three models are trained in a supervised mode (MUSE-S, VM-S, RCLS), whereas only the MUSE and VM models are additionally trained in an unsupervised mode (*model-U*) and a mode that relies on identical strings and numerals (*model-I*).¹

3. Training Data

Most cross-lingual embedding models are rooted in monolingual word embeddings, or they utilise them directly in the training process (Ruder et al., 2019). The prevailing monolingual word embeddings exploited in the majority of papers (Ren et al., 2020; Mohiuddin et al., 2020; Marchisio et al., 2022; Sanigrahi & Read, 2022; etc.) are pre-trained FastText embeddings (Grave et al., 2018).

The FastText embeddings were trained with dimension 300 on the Wikipedia corpus. They are noisy and have limited vocabulary in size and quality. The vocabulary ranges from 300,000 to 800,000 word forms depending on the language and is often polluted with English words or words from other languages. After examining pre-trained Estonian, Slovak, and Korean FastText embeddings, we discovered many English words, for instance, *align*, *score*, *right* (Slovak embeddings), *Norwegian*, *consciousness*, *history* (Estonian embeddings), *and*, *border*, *left* (Korean embeddings) considered as high-frequency words in these languages. The reason for this might not stem only from the data itself but also from the data processing method.

The research in cross-lingual word embeddings has shown that the resulting embeddings' quality heavily depends on the monolingual word embeddings utilised during training (Vulić et al., 2020). Therefore, in this paper, we examine how the different sets of high-quality monolingual word embeddings impact the performance of the models. We compare the common FastText embeddings to the less noisy Sketch Engine embeddings (Herman, 2021). The Sketch Engine embeddings were trained with dimension 100 using the same algorithm as FastText but on a different data

¹ RCLS model is available in the supervised mode only.

source – web corpora, with the vocabulary ranging from two to five million word forms. Table 1 displays the results.

Table 1: Reported precision@1 (P@1) for baseline models trained using FastText (FT) and Sketch Engine (SE) monolingual word embeddings across three language pairs: English-Slovak (en-sk), English-Estonian (en-et), and English-Korean (en-ko)

P@1	en-sk		en-et		en-ko	
	FT	SE	FT	SE	FT	SE
MUSE-S	30.53	37.00	23.87	29.6	25.73	17.47
MUSE-I	29.87	33.40	23.13	24.87	17.61	9.42
MUSE-U	19.07	34.53	0.0	21.73	0.0	0.0
VM-S	35.93	41.20	33.80	32.33	35.29	22.05
VM-I	36.60	36.67	27.67	23.80	22.12	10.92
VM-U	33.13	37.00	25.53	24.93	17.34	10.03
RCLS	38.67	43.40	33.87	38.13	35.63	28.60

When looking at Table 1, we can see that using high-quality monolingual word embeddings, Sketch Engine enhances the performance of all models trained on English-Slovak language pairs and a (slight) majority of the models trained on English-Estonian, within a margin of approximately 0.07% to 15%. The exception is the English-Korean language pair, where models using FastText monolingual word embeddings outperformed all those using Sketch Engine, within a margin of up to 13%.

This demonstrates that not only training data plays a pivotal role but also the quality of the evaluation datasets. In this case, we utilised the MUSE evaluation datasets which, as we show in Section 4, contain many errors, including English-to-English word pairs, predominantly occurring in the English-Korean dataset. This artificially increased the performance of the models trained with FastText monolingual embeddings, which are also heavily polluted with English words. In the following Section, we analyse the evaluation datasets and propose a solution to account for the errors during the evaluation.

4. Evaluation Datasets

The evaluation datasets are compiled based on the vocabulary of the monolingual word embeddings exploited during the training, usually employing an automatic method, such as the popular and widely used MUSE evaluation datasets (Conneau et al., 2018) published along with the eponymous baseline model. These datasets have often been criticised for disproportional part-of-speech distribution, with most word pairs consisting of proper names, which are unsuitable for assessing the quality of the retrieved translation equivalents (Kementchedjieva et al., 2019). However, as our analysis reveals, these datasets contain many other errors.

In this paper, we analyse the errors in three evaluation datasets: English-Slovak, English-Estonian, and English-Korean² and categorise them. Table 2 outlines the

² The analysis of the English-Korean evaluation dataset focuses on the English headwords only.

most common error types occurring in the evaluation datasets. Figures 1-3 illustrate the number of word pairs in each category for each evaluation dataset.

Table 2: Error categories, their description, assigned weight, and an example of a word pair from the respective category

Category	Description	Weight	Example
A	Correct	1	' <i>accidentally</i> ' : ' <i>náhodne</i> ' (sk)
B	Inflected word form	0.8	' <i>workshops</i> ' : ' <i>seminaaride</i> ' (<i>seminaarid</i>) (et)
C	Proper name	0.3	' <i>Ahmed</i> ', ' <i>Bruno</i> ', ' <i>Julia</i> '
D	Part-of-speech mismatch	0.2	' <i>darkness</i> ' : ' <i>temné</i> ' (<i>temnota</i>) (sk)
E	English to English	0.2	' <i>cage</i> ' : ' <i>cage</i> '
F	Abbreviation	0.2	' <i>cnn</i> ', ' <i>fbi</i> ', ' <i>ibm</i> '
G	Synonym	0.7	' <i>customer</i> ' : ' <i>odberateľ</i> ' (<i>zákazník</i>) (sk)
H	Incomplete word	0.3	' <i>cooperative</i> ' : ' <i>ühistu</i> ' (<i>cooperative society</i>) (et)
I	Interjection	0.6	' <i>boom</i> ', ' <i>bang</i> '
J	Missing diacritics	0.4	' <i>everybody</i> ' : ' <i>kazdy</i> ' (<i>každý</i>) (sk)
K	Incorrect	0.1	' <i>lucky</i> ' : ' <i>veab</i> ' (<i>õnnelik</i>) (et)

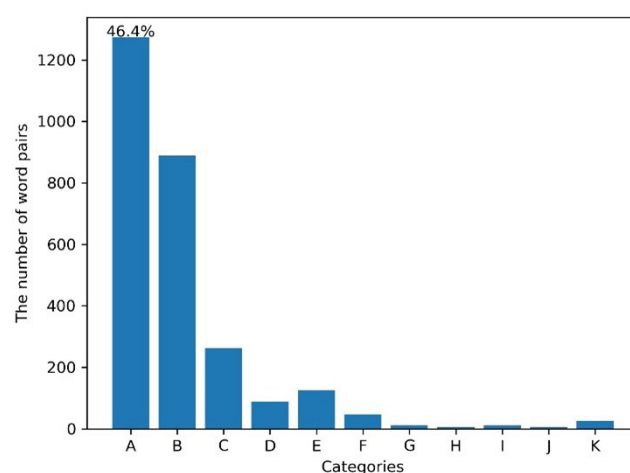


Fig. 1: Number of word pairs in each category in the English-Slovak evaluation dataset

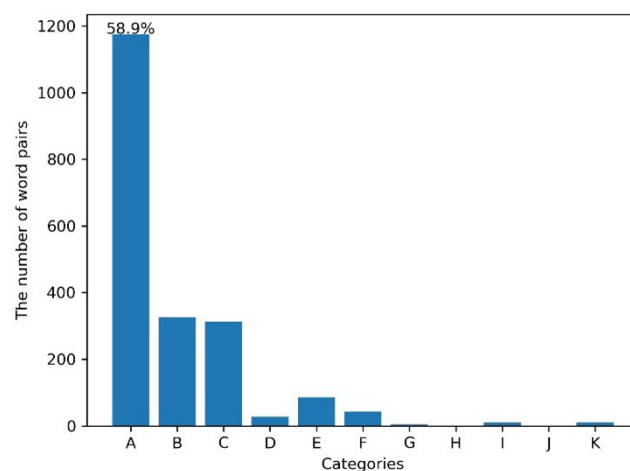


Fig. 2: Number of word pairs in each category in the English-Estonian evaluation dataset

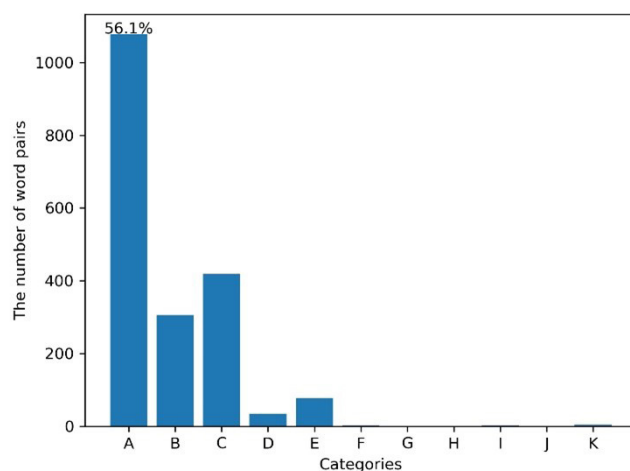


Fig. 3: Number of word pairs in each category in the English-Korean evaluation dataset

According to Table 2, there are 11 types of errors with different levels of severity, such as part-of-speech mismatch (e.g., English-Slovak dataset: *darkness* (noun) – *temné* (adjective, *dark*)), inflected word forms (e.g., English-Estonian dataset: *skill* (sg.) – *oskused* (pl., *skills*) and same-word translations (e.g., English-Slovak dataset: *galaxy* – *galaxy*, English-Korean dataset: *bet* – *bet*), etc. Moreover, given Figures 1–3, we can observe that nearly half of the word pairs in each evaluation dataset contain some error. Such erroneous datasets can distort the actual outcomes.

However, we can allow for these inaccuracies during the evaluation by weighting the precision based on the reliability of the current word pair from the evaluation dataset. In the next step, we assign a weight to each category based on our judgement of the severity of the error, ranging from 1 (the most reliable) to 0.1 (the least reliable). Another reason for weighting is to increase the performance when the model is able to find a word pair with a higher weight and not penalise the model when it does not include a word pair with a low weight. Weights for each category are listed in Table 2. Table 3 shows how the results change when weights are applied.

Table 3: Reported non-weighted (NW) and weighted (W) precision@1 (P@1) for baseline models trained using FastText monolingual word embeddings across three language pairs: English-Slovak (en-sk), English-Estonian (en-et), and English-Korean (en-ko)

P@1	en-sk		en-et		en-ko	
	NW	W	NW	W	NW	W
MUSE-S	30.53	26.09	23.87	21.61	25.73	18.81
MUSE-I	29.87	24.28	23.13	17.69	17.61	12.82
MUSE-U	19.07	16.34	0.0	0.0	0.0	0.0
VM-S	35.93	29.52	33.80	28.73	35.29	27.11
VM-I	36.60	29.25	27.67	22.61	22.12	15.90
VM-U	33.13	28.11	25.53	21.26	17.34	13.69
RCLS	38.67	33.28	33.87	30.13	35.63	27.86

Based on Table 3, when weights are applied, the precision decreases by around 2% to 8%, while the gaps are closer to the upper limit for the English-Korean language combination. This signifies that the English-Korean evaluation dataset contained more severe error types, and the models were better at finding such word pairs.

On top of that, adjusting the weights altered the rankings of the models' performance. For example, the VM-I model, evaluated on the English-Slovak language pair, initially outperformed the VM-S model when no weights were applied. However, upon introducing weights, the VM-S model demonstrated superior performance.

5. Evaluation Metric

Another problematic aspect of the evaluation is the employed metric. The most reported metric is precision at k ($P@k$), where k represents the number of translation-equivalent candidates retrieved for a headword. In many papers, the k is set to 1, 3 or 5, i.e., fixed. However, the number of translation equivalents for each headword differs from word to word and dataset, including the MUSE evaluation datasets. Table 4 shows the statistics of the number of translation equivalents in the English-Slovak (en-sk), English-Estonian (en-et), and English-Korean (en-ko) evaluation datasets.

Table 4: Number of translation equivalents (TEs) in three MUSE evaluation datasets

TEs	en-sk	en-et	en-ko
1	760	1111	1085
2	395	303	310
3	208	71	63
4	88	14	7
5+	40	1	0

Table 4 confirms that having a fixed k is not an appropriate setting for an evaluation. Additionally, reporting purely on precision is not informative about the actual model's performance. Denisová & Rychlý (2023) showed that the higher the precision the model obtains, the lower the recall. This is also caused by fixed k , which often imbalances precision and recall. Instead of preferring one or the other, we can implement a simple classification neural network³ which classifies the word pairs found by the cross-lingual embedding models as correct or incorrect while allowing for a dynamic k . The aim is to identify as many relevant translation equivalents as possible for each headword and report precision, recall, and F1 scores not constrained by a predefined set of k translation-equivalent candidates while balancing precision and recall and maintaining high performance. Tables 5, 6, and 7 outline the results of precision, recall, and F1 scores evaluated using fixed and dynamic k .

³ The technical details of the classification neural network and an extensive evaluation across a wide range of language pairs will be published in a separate paper.

Table 5: Reported precision (P), recall (R), and F1 score using fixed k ($k = 1$) and dynamic k (NN) for the baseline model VM-S trained using FastText monolingual word embeddings across three language pairs: English-Slovak (en-sk), English-Estonian (en-et), and English-Korean (en-ko)

VM-S	P@1	P@NN	R@1	R@NN	F1@1	F1@NN
en-sk	35.93	34.83	19.68	44.30	25.43	39.00
en-et	33.80	36.13	25.46	42.16	29.05	38.91
en-ko	35.29	39.06	26.90	45.45	30.53	42.02

Table 6: Reported F1 scores using fixed k ($k = 1$) and dynamic k (NN) for baseline models trained using FastText monolingual word embeddings across three language pairs: English-Slovak (en-sk), English-Estonian (en-et), and English-Korean (en-ko)

	en-sk		en-et		en-ko	
	F1@1	F1@NN	F1@1	F1@NN	F1@1	F1@NN
MUSE-S	21.61	32.14	20.51	20.18	22.26	30.86
MUSE-I	21.14	29.59	19.88	33.03	15.23	15.22
MUSE-U	13.49	34.44	0.0	0.0	0.0	0.0
VM-S	25.43	39.00	29.05	38.91	30.53	42.02
VM-I	25.90	33.00	23.78	30.77	19.13	18.54
VM-U	23.45	41.72	21.94	32.77	15.00	22.05
RCLS	27.36	36.02	29.10	33.03	30.82	41.07

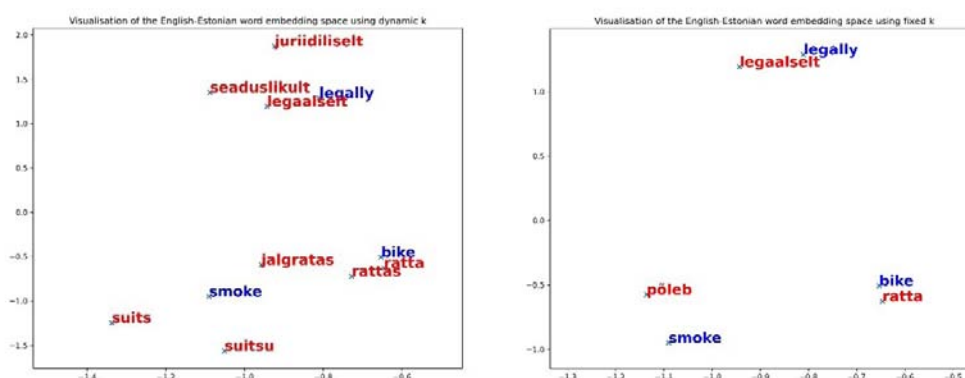
Table 7: Reported F1 scores using fixed k ($k = 1$) and dynamic k (NN) for baseline models trained using Sketch Engine monolingual word embeddings across three language pairs: English-Slovak (en-sk), English-Estonian (en-et), and English-Korean (en-ko)

	en-sk		en-et		en-ko	
	F1@1	F1@NN	F1@1	F1@NN	F1@1	F1@NN
MUSE-S	26.19	45.11	25.44	40.00	15.12	23.62
MUSE-I	23.64	43.92	21.37	29.17	8.15	20.83
MUSE-U	24.44	41.44	18.68	32.89	0.0	0.0
VM-S	29.16	47.59	27.79	39.63	19.07	33.09
VM-I	25.95	42.76	20.45	27.81	9.45	18.93
VM-U	26.19	46.98	21.43	36.97	8.68	13.79
RCLS	30.71	51.74	32.77	44.55	24.74	40.51

Given Tables 5, 6, and 7, we can observe that a dynamic k significantly increases F1 scores, in most cases within a margin of approximately 0.20% to 30%, offering a balanced trade-off between precision and recall. Although P@1 sometimes yields better results, it only assesses a small part of the models' performance. Employing the F1 score in the evaluation process provides a more accurate picture of the model's performance and balances the precision and recall. Table 8 and Figure 4 demonstrate an example of how dynamic and fixed k work.

Table 8: Examples from the model VM-S trained using FastText monolingual word embeddings on the English-Estonian language pair

Evaluation dataset		Dynamic k	k	Fixed $k = 1$
legally	juriidiliselt	✓	3	-
	legaalselt	✓		✓
	seaduslikult	✓		-
	õiguslikult	-		-
smoke	suitsu	✓	2	-
	suits	✓		-
	-	-		põleb
bike	rattas	✓	3	-
	ratta	✓		✓
	jalgratas	✓		-

**Fig. 4:** Visualisation of the English-Estonian word embedding space using dynamic (left) and fixed (right) k

When looking at Table 8 and Figure 4, we can see that setting a fixed k significantly limits the cross-lingual embedding space by providing only one translation-equivalent candidate with the closest embedding. On the other hand, the classification neural network enables us to set k for each headword individually. For example, the headword ‘legally’ corresponds to four translation equivalents: ‘juriidiliselt’, ‘legaalselt’, ‘seaduslikult’, and ‘õiguslikult’. When using a fixed $k = 1$, the model identified only the closest embedding equivalent, ‘legaalselt’. However, with a dynamic k , the model successfully retrieved three of the translation equivalents, i.e., set $k = 3$.

Another interesting example is the incorrectly selected translation-equivalent candidate ‘põleb’ by the model using fixed k , which did not occur in the evaluation dataset. In the cross-lingual embedding space, ‘põleb’ was the closest match. Thus, with $k = 1$ (fixed), the model identified it as the only candidate. Conversely, a dynamic k allowed the model to choose candidates whose embeddings were further from the headword’s embedding, such as ‘suits’ and ‘suitsu’ from the evaluation dataset.

6. Conclusion

In this paper, we have introduced cross-lingual embedding models for lexicography, showing the advantages and limitations of certain types of training data, evaluation datasets, and employed metrics. We have demonstrated through extensive analysis across three language pairs that the quality of both training and evaluation data significantly influences the performance of these models. Despite the promising capabilities of models like MUSE, VecMap, and RCLS in bridging linguistic gaps, our findings reveal that enhancements in data quality yield more substantial improvements than algorithmic optimisation alone.

Our investigation into various types of training data underscores the necessity for employing less noisy, more comprehensive monolingual word embeddings such as those from Sketch Engine over commonly used sources like FastText. Similarly, our analysis of evaluation datasets has shown substantial errors and biases that can skew the models' performance. By implementing weighted precision measures that account for the reliability of word pairs in the evaluation datasets, we have provided a more nuanced view of model effectiveness.

Furthermore, the shift from a fixed to a dynamic *k*-based evaluation approach has proven essential in capturing a more accurate representation of model capabilities, as it allows for a balance between precision and recall.

Reformulated, and answering the question posed in the subtitle of our paper, while cross-lingual embedding models are useful tools for computational lexicography, their advancement depends crucially, first, on the quality of the training data, over and above mere algorithmic optimisation. Second, given that not all word pairs in the evaluation datasets have the same reliability, one must and *can* reflect that using weights for the evaluation. Third, given that words often have multiple translation equivalents across languages, one must and *can* also reflect that variation in the model, thereby significantly improving F1 scores.

Future research should continue to focus on refining data quality and developing more robust evaluation methods to fully harness the potential of these models in the field of lexicography. Our work serves as a foundational step towards improving the reliability and efficacy of translation-equivalent retrieval, aiming to support lexicographers in their bilingual dictionary projects with more accurate and resource-efficient tools. Additionally, it is crucial to engage lexicographers in the development process to ensure that these computational advances are practically implementable and truly beneficial in real-world lexicographic practice, addressing the common challenge of bridging theoretical usefulness with practical application.

References

Artetxe, M., Labaka, G., & Agirre, E. (2018a). A Robust Self-learning Method for Fully Unsupervised Cross-lingual Mappings of Word Embeddings. In I. Gurevych, & M. Yuseke

(Eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 789–798). Association for Computational Linguistics.

Artetxe, M., Labaka, G., & Agirre, E. (2018b). Generalizing and Improving Bilingual Word Embedding Mappings with a Multi-Step Framework of Linear Transformations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), 5012–5019. <https://doi.org/10.1609/aaai.v32i1.11992>

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., & J'egou, H. (2018). Word Translation Without Parallel Data. In *Proceedings of International Conference on Learning Representations* (pp. 1–13).

Denisová, M., & Rychlý, P. (2023). Evaluation of the Cross-lingual Embedding Models from the Lexicographic Perspective. In M. Medveď, M. Měchura, C. Tiberius, I. Kosem, J. Kallas, M. Jakubiček, & S. Krek (Eds.), *Electronic lexicography in the 21st century (eLex 2023): Invisible Lexicography. Proceedings of the eLex 2023 conference* (pp. 1–18). Lexical Computing CZ, s.r.o.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). Learning Word Vectors for 157 Languages. In N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, & T. Tokunaga (Eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (pp. 3483–3487). European Language Resources Association (ELRA).

Herman, O. (2021). Precomputed Embeddings for 15+ Languages. In A. Horák, P. Rychlý, & A. Rambousek (Eds.), *Proceedings of the Fifteenth Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2021* (pp. 41–46). Tribun EU.

Joulin, A., Bojanowski, P., Mikolov, T., Jégou, H., & Grave, E. (2018). Loss in Translation: Learning Bilingual Word Mapping with a Retrieval Criterion. In E. Riloff, D. Chiang, J. Hockenmaier, & J. Tsujii (Eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 2979–2984). Association for Computational Linguistics.

Kementchedjhieva, Y., Hartmann, M., & Søgaard, A. (2019). Lost in Evaluation: Misleading Benchmarks for Bilingual Dictionary Induction. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 3336–3341). Association for Computational Linguistics.

Marchisio, K., Saad-Eldin, A., Duh, K., Carey P., & Koehn, P. (2022). Bilingual Lexicon Induction for Low-Resource Languages using Graph Matching via Optimal Transport. In Y. Goldberg, Z. Kozareva, & Y. Zhang (Eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 2545–2561). Association for Computational Linguistics.

Mohiuddin, T., Bari, S. M., & Joty, S. (2020). LNMap: Departures from Isomorphic Assumption in Bilingual Lexicon Induction Through Non-Linear Mapping in Latent Space. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 2712–2723). Association for Computational Linguistics.

Ren, S., Liu, S., Zhou, M., & Ma, S. (2020). A Graph-based Coarse-to-fine Method for Unsupervised Bilingual Lexicon Induction. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault

(Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 3476–3485). Association for Computational Linguistics.

Ruder, S., Vulić, I., & Søgaard, A. (2019). A Survey of Cross-lingual Word Embedding Models. *The Journal of Artificial Intelligence Research*, 65, 569–631. <https://doi.org/10.1613/jair.1.11640>

Sannigrahi, S., & Read, J. (2022). Isomorphic Cross-lingual Embeddings for Low-Resource Languages. In *Proceedings of the 7th Workshop on Representation Learning for NLP (RepL4NLP)* (pp. 133–142). Association for Computational Linguistics.

Smith, N. A. (2019). Contextual Word Representations: A Contextual Introduction. ArXiv, abs/1902.06006.

Vulić, I., Korhonen, A., & Glavaš, G. (2020). Improving Bilingual Lexicon Induction with Unsupervised Post-Processing of Monolingual Word Vector Spaces. In S. Gella, J. Welbl, M. Rei, F. Petroni, P. Lewis, E. Strubell, M. Seo, & H. Hajishirzi (Eds.), *Proceedings of the 5th Workshop on Representation Learning for NLP* (pp. 45–54). Association for Computational Linguistics.

Websites:

Sketch Engine Embeddings. Retrieved May 5, 2024, from <https://embeddings.sketchengine.eu/>

FastText. Retrieved May 5, 2024, from <https://fasttext.cc/>

Wikipedia. Retrieved May 5, 2024, from <https://www.wikipedia.org/>

Contact information

Michaela Denisová

Natural Language Processing Centre, Faculty of Informatics, Masaryk University
449884@mail.muni.cz

Gilles-Maurice de Schryver

Ghent University (BantUGent & LT³) & University of Pretoria (African Languages)
gillesmaurice.deschryver@UGent.be

Pavel Rychlý

Natural Language Processing Centre, Faculty of Informatics, Masaryk University & Lexical Computing
pary@fi.muni.cz

Stephanie Evert, Christine Ganslmayer, and Christian Rink

MULTI-LEVEL ANALYSIS AS A SYSTEMATIC APPROACH TO EVALUATING THE QUALITY OF AI-GENERATED DICTIONARY ENTRIES

Abstract In this paper, we report on our development of a multi-level analysis framework that allows us to assess AI-generated lexicographic texts on both a quantitative and qualitative level and compare them with human-written texts. We approach this problem through a systematic and fine-grained evaluation, using dictionary articles created by human subjects with the help of ChatGPT as an example. The levels of our framework concern the assessment of individual entries, a comparison with existing dictionary entries written by experts, an analysis of the writing experiment, and the discussion of AI-specific aspects. For the first level, we propose an elaborate evaluation grid that enables a fine-grained comparison of dictionary entries. While this grid has been developed for a specific writing experiment, it can be adapted by metalexicographical experts for the evaluation of all kinds of dictionary entries and all kinds of dictionary information categories.

Keywords Conversational AI; generated lexicographical data; dictionary criticism; evaluation methodology; semantic analysis; standard dictionaries

1. Introduction: LLMs and Dictionary Criticism

After the November 2022 release of ChatGPT, a freely accessible and easy-to-use conversational AI, text-generating AI products based on large language models (LLMs) have quickly become a part of daily life. The resulting transformation of communicative practice will have a profound impact on all areas of text production and reception, as well as knowledge society more broadly. This includes, of course, the field of lexicography, and several pilot studies have already explored possible applications of ChatGPT in lexicography; de Schryver (2023, p. 377) concludes:

We have seen that, with the right prompts, an LLM like ChatGPT can already be brought in to either compile a dictionary on its own, or, somewhat more safely, to speed up dictionary compilation by providing quality draft material which human lexicographers then assess and improve upon.

While most authors eventually come to the conclusion that ChatGPT cannot replace professional lexicographical work yet (e.g., Arias-Arias et al., 2024), its results are often astonishingly good at first glance, depending on the target language and the particular version of ChatGPT used. This impression was confirmed by an exploratory case study we carried out: Laypersons without prior lexicographical knowledge were not able to tell reliably whether a dictionary entry for a monolingual German general dictionary had been created by a human or by a human-guided AI. Reassuringly,

lexicographical experts performed much better at this task. In a similar vein, Lew (2023, pp. 8–9) finds that ChatGPT is at least partially capable of generating quality dictionary entries: whereas definitions seem to be “indistinguishable in quality from those written by highly trained human lexicographers”, “examples [...] turned out not to be as impressive, and significantly worse than those crafted by professional human lexicographers.”

This raises many questions, and the consequences for dictionaries as products, for their planning and use as well as their critical evaluation call for metalexicographical reflection (cf. Wiegand, 1998, pp. 79–80). Most people still have a very limited understanding of what LLMs are and what they can and cannot do. Our paper aims to shed some light on how good conversational AI has already become at (co-)authoring dictionary entries. Following de Schryver (2023, p. 377), we focus on a semi-automatic approach where AI-generated dictionary entries are reviewed and improved by human writers. The foremost and crucial question in this endeavour is how to achieve a detailed and comprehensive evaluation of the quality of dictionary entries. For this purpose, we developed a multi-level analysis framework for AI-generated dictionary entries and tested it in a writing experiment with non-expert writers assisted by AI.

Various approaches have been proposed in the field of dictionary criticism: Svensén (2009, pp. 482–485) describes several procedures for a dictionary evaluation: 1. analysis of the dictionary by one or more reviewers, 2. dictionary comparison, 3. use of dictionaries by test subjects with subsequent interview, 4. specially designed usage situations with specific user groups such as language learners. Tarp (2017) lists various criteria that relate to the functional type of a dictionary and its target users (see also Engelberg & Lemnitzer, 2009, pp. 190–222; Nielsen, 2009; Pearsons & Nichols, 2013). Further evaluation criteria include the quantity and quality of lexicographical information, their presentation, and the design of entries (cf. Tarp, 2017, p. 127). It is highly desirable for qualitative aspects to play a prominent role in the evaluation, even though it is more difficult to operationalise them than quantitative criteria.

Such criteria have to be developed into a concrete evaluation framework that can be applied successfully and adapted to different use cases. Dictionary criticism research usually applies ordinal scales for the evaluation of each aspect. Ripfel’s (1989, p. 57) framework, for example, is based on an analysis of 736 dictionary reviews, from which she derived evaluation categories (e.g., quality of definitions) that are to be judged on a five-point scale (e.g., from “easily comprehensible/simple” to “incomprehensible/cumbersome”). Pearson & Nichols (2013) also propose a five-point scale, but address only quantitative aspects. Lew (2023) asked four experts to evaluate dictionary entries for 15 communication verbs generated by GPT-4, focusing on only three information categories (definitions, examples, entry as a whole) and rating them on a generic five-point scale (from bad to great). However, Lew does not provide precise evaluation criteria.

In our opinion, the evaluation of dictionary entries can be improved substantially if we adopt results and insights on text quality from German text linguistics (Nussbaumer, 1991; Sieber & Nussbaumer, 1994; Fritz, 2017; Abel et al., 2020), empirical writing

didactics (Becker-Mrotzek, 2014) and qualitative media research and pedagogy (Mayring & Hurst, 2017). These fields consider an evaluation to be more reliable if, according to the analytical method, individual characteristics of text quality are determined according to text type, and then transformed into an evaluation grid of specific criteria to be judged. The overall evaluation is then based on awarding points for each criterion (cf. Neumann, 2017, pp. 208–211; Mayring & Hurst, 2017, p. 498; Becker-Mrotzek, 2014, pp. 507–510). Our paper outlines the development of such an evaluation grid for entries in a monolingual (German) general dictionary (Section 4). It is based on a general multi-level analysis framework (Section 3), which assesses texts (here: dictionary entries) according to various quantitative and qualitative criteria in comparison to existing reference texts of the same type. Our evaluation grid can be adapted to other functional types of dictionaries and is ideally suited for the evaluation of AI-generated dictionary entries, highlighting the specific strengths, weaknesses, and challenges of AI-assisted text production. Both the framework and the evaluation grid have been developed and tested in the context of a writing experiment, which is briefly described in Section 2. The application of our framework in Section 5 also specifically addresses the question how the specific limitations and challenges of text generation with an AI chatbot can be worked out and potential solutions can be sought. The flowchart in Fig. 1 gives a visual overview of the overall approach and methodology.

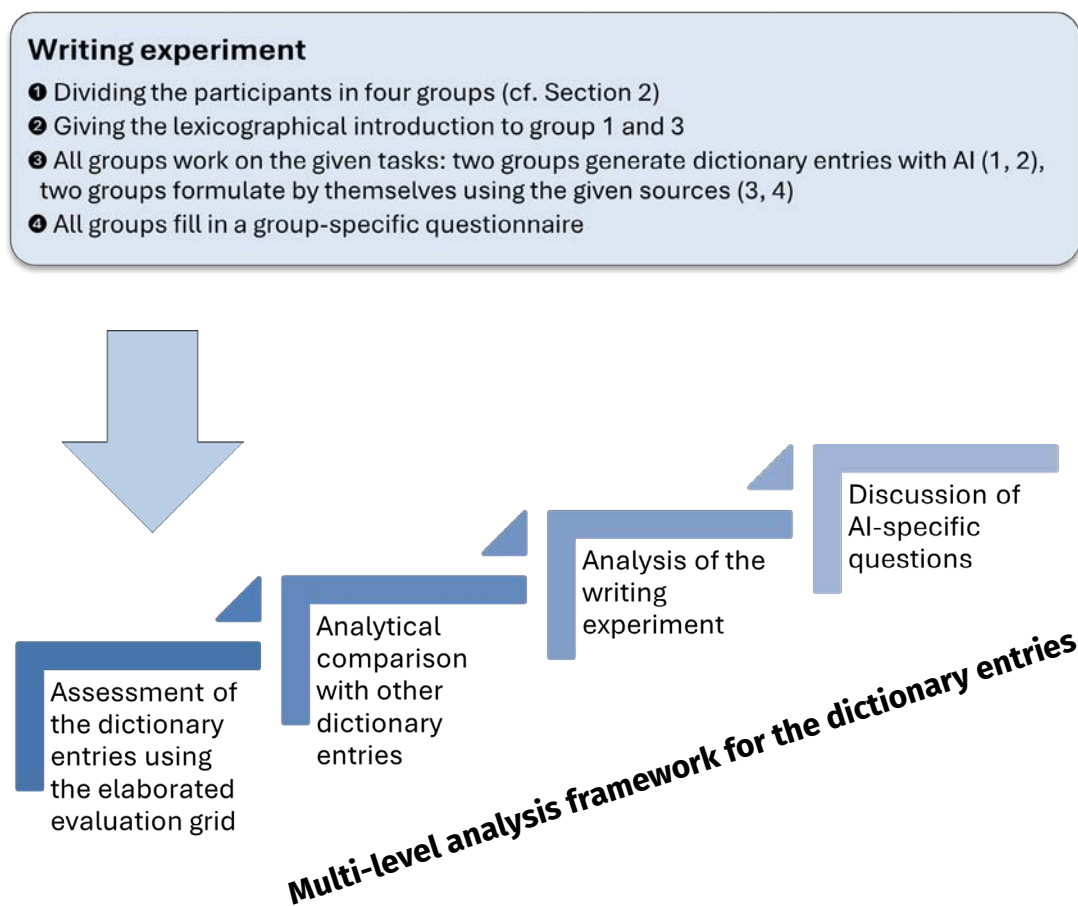


Fig. 1: Overall approach and methodology

2. Writing Experiment on Dictionary Entries

Our non-representative exploratory writing experiment was carried out in April 2024 in an advanced seminar in German linguistics at FAU Erlangen-Nürnberg. The task was to write dictionary entries with different levels of prior knowledge of the text type and with or without support from conversational AI. Six of the ten student participants (mainly from German and English studies) were asked to generate dictionary entries using ChatGPT (GPT-3.5), the other four were asked to write the entries independently without computational aid. The experiment was conducted in three phases, as outlined below: (1) To introduce prior knowledge of the text type as an independent variable, half of the students in each group were given a brief lexicographical introduction (15 minutes), presenting the components of dictionary microstructure and illustrating them with examples (entry *Pferd* [horse] in *Duden Bedeutungswörterbuch* [Duden Explanatory Dictionary] and *DWDS* [Digital Dictionary of the German Language]). We also introduced the modern corpus-based approach to lexicography, which integrates authentic evidence into the entries. (2) The instructions for the writing phase of 45 minutes were as follows:

Groups with AI:

Create dictionary entries for the lemmata *Maus* [mouse] and *köpfen* [behead]. Use ChatGPT 3.5 and document all your attempts by copying both your input and the results (chatbot output) into a Word document (in chronological order). Only use the AI assistant. Internet sources must not be used under any circumstances so as not to distort the results of the study.

Groups without AI:

Create dictionary entries in Word for the lemmata *Maus* [mouse] and *köpfen* [behead]. Use the provided examples as source material. AI and other Internet sources must not be used under any circumstances so as not to distort the results of the study.

Groups without AI were given six sheets of example sentences for each lemma to simulate the lexicographical process of identifying different word senses and sub-senses, and to support the semantic analysis. Groups with AI had to rely on the AI training data exclusively. (3) After the writing phase, participants filled in questionnaires about their background and experiences from the experiment (20 minutes). A detailed evaluation of these questionnaires is beyond the scope of the present paper.

3. A Multi-Level Analysis Framework for AI-generated Dictionary Entries

Our multi-level analysis framework for AI-generated dictionary entries builds on and extends prior work by several authors: The general areas of the evaluation framework are inspired by Svensén (2009). Our evaluation categories for an entry in a monolingual German general dictionary are based on the well-known parameters

of dictionary microstructure, in combination with the evaluation criteria of Nielsen (2009) and Tarp (2017). The evaluation grid takes up aspects of Ripfel (1989) and Pearsons & Nichols (2013). Evaluation categories and criteria were tested empirically on dictionary articles (e.g., from the writing experiment) and refined if necessary.

Our aim is to develop an analysis framework that can be adapted by experts for the evaluation of different types of (AI-generated) dictionaries, but can also be applied more widely to assess AI-generated dictionary entries and similar texts without having to resort to simplistic automated procedures (e.g., Celikyilmaz et al., 2020).

For the evaluation of an individual dictionary entry, only aspects of the microstructure, target user requirements, and dictionary type are relevant; other metalexicographical categories such as macro, medio or access structure do not apply. A strong focus was placed on semantic categories, complemented by general qualitative aspects such as aesthetic, content-related and linguistic appropriateness of the text, relative to its text type. An additional point of reference was provided by sample entries from existing dictionaries. The resulting categories and criteria form the basis of a finely differentiated evaluation grid (cf. Section 4), which enables judges to award points to each individual criterion.

Since we agree with de Schryver (2023, p. 377) that AI-generated dictionary entries will in most cases have to be checked and improved by humans, a well-founded evaluation of such entries can only be carried out in the context of a writing experiment that includes teams of AI and human as test subjects. The analysis of this experiment also has to address AI-specific questions, e.g., what influence prompts provided to the AI have on the quality of the output.

Our complete systematic approach to the evaluation of AI-generated dictionary entries thus encompasses four levels of increasing breadth and generality.

3.1 Assessment of Dictionary Entries via an Elaborate Evaluation Grid

The first level of our multi-level analysis framework evaluates the AI-generated dictionary entries using an elaborate evaluation grid, which contains quantitative and qualitative criteria for every information category in the form comment and semantic comment of a dictionary entry (such as pronunciation indication, polysemy, or examples). Our evaluation grid and the methodology for its development are presented in detail in Section 4. The grid can be applied equally well to dictionary entries in printed and online dictionaries written by human experts, provided that the necessary adaptations are made to align with the specific dictionary type.

3.2 Comparison With Other Dictionary Entries

The second level of our analysis framework mandates a comparison with other dictionary entries. In our case, entries from writing experiments (cf. Section 3.3) are

evaluated according to Section 3.1 and compared to each other. Additional comparisons are made with existing dictionary entries written by professional lexicographers (here: from *duden.de*). These entries have been written by experts and should therefore be recognised as a reference, even if they do not achieve a perfect score according to our grid (e.g., because collocations are missing or uncommon (sub-)senses are included; see Section 4.3 for details).

3.3 Analysis of Writing Experiments

The next level of the evaluation framework calls for one or more writing experiments to be carried out and analysed. In most cases, test subjects are humans with different levels of expertise, as well as teams of AI and human. The analysis could focus on aspects such as comparing dictionary entries written with and without AI support, the effect of human post-editing on AI-generated dictionary entries, the test subjects' previous experience with dictionaries and AI tools, etc. Of course, writing experiments that target human-written texts exclusively are also possible.

3.4 AI-specific Considerations

Since our evaluation framework has been developed with AI-generated dictionary entries in mind, AI-specific questions should be considered in the fourth level. Typical examples include: (1) information categories that the AI cannot generate, (2) the prompt used to instruct the AI, (3) comparison of writing experiments with purely AI-generated dictionary entries, using a prompt optimised by a lexicographic expert, and (4) the discussion of AI-specific errors and deficiencies.

4. Developing an Elaborate Evaluation Grid

For the evaluative assessment of a dictionary entry, the entire entry and the lexicographical information provided are analysed in regard to the two main aspects of (a) quantity and (b) quality. Quantity refers to the presence of information as well as its scope and completeness. Quality refers to the accuracy and appropriateness of both content and language, as well as the authenticity, etc. of the information provided. Specific qualitative and quantitative criteria are established for each information category. Our main focus (partly depending on dictionary type) is on the qualitative criteria, which are thus more numerous and can be awarded more points. First of all, the targeted dictionary type has to be identified, indicating a possible information programme (*sensu* Engelberg & Lemnitzer, 2009, p. 25) of the microstructure, which in turn determines the information categories to be evaluated. Our dictionary comparison (according to Section 3.2) uses lemma *Maus* [mouse] from the online dictionary *duden.de*, which claims to offer “umfassende Informationen zu Rechtschreibung, Grammatik und Bedeutung eines Wortes” [comprehensive information on spelling, grammar, and meaning of a word] and “den richtigen Gebrauch sowie die Aussprache und Herkunft eines Wortes und [...] dessen Synonyme” [correct usage, pronunciation and origin of a word and [...] its synonyms]

(duden.de).¹ Dictionaries with this amount of information are referred to as poly-informative dictionaries (*sensu* Wiegand, 2010, p. 85), which belong to the class of general dictionaries (*sensu* Engelberg & Storrer, 2016, p. 41) because they do not focus on specific information (unlike e.g., a dictionary of synonyms) and they are not limited in the selection of lemmas (unlike e.g., a dictionary of loanwords). Table 1 shows the main sections and information categories (“Angaben” *sensu* Wiegand, 1989, p. 468) to be analysed.

Table 1: Main sections + information categories of our evaluation grid

Overall microstructure	text structure / design, overall assessment
Comment on form	lemma, pronunciation indication, grammar indication, orthography indication, frequency
Comment on semantics	polysemy, pragmatics, meaning, examples, collocations, word formation, proverbs & idioms, synonymy, etymology

Specific quantitative and qualitative criteria then have to be developed individually for each category. For the overall structure and comment on form, they are mainly based on metalexicographical categories, from which specific questions can in turn be derived (cf. Tarp, 2017, pp. 122–125). The main focus of our evaluation is on the different information categories in the comment on semantics. The development of corresponding criteria is exemplified below for the category of meaning indication.

4.1 Exemplary Derivation of Quantitative and Qualitative Criteria

We illustrate the derivation of criteria for the evaluation grid on the example of the information category meaning indication. Quantitative criteria refer only to its presence and completeness, i.e., whether it is included at all and whether a definition has been provided for each (sub-)sense. Quantitative criteria are each awarded 0 or 1 point.

Qualitative criteria relate to use of standard language, redundancy, genus proximum, differentia specifica, linguistic correctness, accuracy of content and information density. Use of standard language requires meaning indications to be formulated without specialised terminology or, e.g., orientation towards a particular sociolect; redundancy should generally be avoided; the integration of genus proximum and differentia specifica aims to systematically differentiate words and position them within a semantic category; linguistic correctness is satisfied if there are no orthographic or grammatical errors; accuracy of content requires that the provided meaning indication is correct and complete; and information density refers to the length of the meaning indication, e.g., whether only a single word (meaning equivalent) is given or a complete semantic paraphrase. Qualitative criteria can be awarded up to 2 points, which gives them more weight in the overall evaluation and allows for more fine-grained judgments (awarding 1 point if a criterion is only partially satisfied).

¹ <https://www.duden.de/woerterbuch> (retrieved May 31st, 2024)

4.2 Evaluation Grid

Table 2 shows the complete evaluation grid with all categories and individual criteria. There is no room here for a full derivation of the evaluation criteria (as exemplified in Section 4.1), which will be detailed in a follow-up publication. In addition to providing a framework for evaluating the results of the writing experiment (Section 2), our evaluation grid is to be understood as a metalexicographical contribution to the systematic evaluation of dictionary entries. The grid can (and needs to) be adapted for application to different types of dictionaries: in bilingual dictionaries, e.g., other information categories are relevant and other criteria have to be formulated (less differentiation of the meaning indication, but inclusion of translation equivalents, greater importance of collocation information, etc.).

Table 2: Evaluation grid for the microstructure of a dictionary entry

Categories (Evaluation areas)	Criteria (max. 136)	Quantity (0–1 each, total max. 30)
		Quality (0–2 each, total max. 106)
Whole microstructure (max. 18)		
Textual structure / design (max. 8)	Non/typographical microstructural indicator, paragraphs, entry structure recognisable	
	Correctness microstructural indicator, understandability	
Overall rating (max. 10)	Dictionary functions fulfilled, target users considered, conciseness of content, linguistic conciseness, use of general language	
Comment on form (max. 32)		
Lemma (max. 3)	Availability	
	Linguistic correctness	
Pronunciation (max. 6)	Availability, accentuation available	
	Correctness, accentuation correct	
Grammar (max. 14)	Indication of part of speech available, genus specification available, inflection available, paradigm fully covered	
	Part of speech correct, genus indication correct, inflection indication correct, no redundancy, paradigm correct	
Orthography (max. 6)	Availability, hyphenation available	
	Correctness, hyphenation correct	
Frequency (max. 3)	Availability	
	Correctness	
Comment on semantics (max. 86)		
Polysemy (max. 4)	Availability, completeness	
	Usualness	
Pragmatics (max. 6)	Availability, completeness	
	Linguistic correctness, differentiation	
Meanings (max. 16)	Availability, completeness	
	Use of general language, information density, linguistic correctness, accuracy of content, genus proximum, differentia specifica, no redundancy	

Examples (max. 14)	Availability, completeness
	Clarity, linguistic correctness, accuracy of content, authenticity, appropriateness of content, sources provided
Collocations (max. 14)	Availability, amount
	Clarity, linguistic correctness, accuracy of content, authenticity, empirically verifiable, aligned with meanings
Proverbs & idioms (max. 5)	Availability
	Linguistic correctness, authenticity
Word formation (max. 10)	Availability, amount
	Linguistic correctness, accuracy of content, multiple types of word formation covered, usualness
Synonymy (max. 10)	Availability, completeness
	Linguistic correctness, content accurate, usualness, empirically verifiable
Etymology (max. 7)	Availability
	Linguistic correctness, accuracy of content, information density

The evaluation grid in Table 2 comprises a total of 136 points to be awarded, with 30 points for quantitative criteria and 106 points for qualitative criteria. A maximum of 18 points are awarded for general categories, 32 points for the comment on form, and 86 points for comment on semantics. For the overall evaluation of an entry, the total number of points is converted to a percentage and graded based on the ranges shown in Table 3, with 94–100% representing an excellent result (grade I) and anything below 50% considered inadequate for a dictionary (grade V).

Table 3: Grading scheme for the overall evaluation

98–100%	133–136 points	Grade I: <i>meets quantitative and qualitative requirements for the dictionary type excellently</i>
94–97%	128–132 points	
90–93%	122–127 points	Grade II: <i>meets quantitative and qualitative requirements for the dictionary type well (above par)</i>
85–89%	116–121 points	
80–84%	109–115 points	
76–79%	103–108 points	Grade III: <i>meets quantitative and qualitative requirements for the dictionary type satisfactorily</i>
71–75%	97–104 points	
66–70%	90–96 points	
61–65%	83–89 points	Grade IV: <i>meets quantitative and qualitative requirements for the dictionary type fairly (below par)</i>
50–60%	68–82 points	
0–49%	00–67 points	Grade V: <i>does not meet quantitative and qualitative requirements for the dictionary type</i>

4.3 Exemplary Assessment of the Lemma *Maus* (duden.de)

As an example for the application of our evaluation grid, we briefly describe the assessment of the lemma *Maus* [mouse] on duden.de; the result also serves as a point of reference for the dictionary comparison in the second level of our evaluation

framework. This dictionary entry achieves an overall rating of 70% with only 95 out of 136 points and therefore has to be located in the lower range of Grade III. Of course, this rating cannot be generalised to the dictionary as a whole. For example, the lemma *Katze* [cat] as a control entry is at least in the upper range of Grade III (with 78%).

Due to space constraints, we can only describe the assessment of the category Meanings in detail. Meaning indications are provided in the *duden.de* entry on *Maus*, but are incomplete for sense 6 “weibliche Scham; Vulva” [female pubis, vulva] (lacking even a link to the lemma *Vulva*, as is provided for sense 2 “Kosewort” [term of endearment]). The meaning indications or links provided are correct in terms of language and content. They use general language, being neither technical nor colloquial. A classification of the lemma into superordinate semantic categories such as biological order is also provided: “kleines [graues] Nagetier” [small [grey] rodent], although differentiation within the category is too unspecific. The two sub-senses “4. a) Geld [money]” and “b) Euro, Mark o.Ä. [euro, mark, etc.]” appear to be redundant and over-differentiated. Information density varies: while meaning indications for senses 1 (rodent) and 5 (computer mouse) are relatively extensive, others such as sense 6 (female pubis, vulva) are only paraphrased with a synonymous word group. Usualness is evaluated in the polysemy category because it may differ between (sub-)senses: it is at least questionable whether sense 6 is indeed frequently used in general language.²

The poor overall score of the dictionary entry is partly due to the absence of nominal compounds (in the word formation category) and collocations, losing a total of 24 possible points. It is thus a crucial advantage of our evaluation grid that dictionary entries can be judged at a more fine-grained level (as will be shown in Table 4).

5. Application of the Multi-level Analysis Framework

In this section, we evaluate the dictionary entries from our writing experiment according to the multi-level analysis framework presented in Section 3.

5.1 Assessment of the Dictionary Entries

The dictionary entries from the writing experiment were independently evaluated by two experts on the basis of the evaluation grid shown in Table 2. As an added benefit, the grid allows us to validate the reliability of ratings much better than by comparing overall scores.³ For this purpose, we calculate Cohen’s kappa as a measure of the inter-annotator agreement between the two experts (Artstein & Poesio, 2008) and thus the reliability of the assessment. Kappa values usually range from 0 to 1, with scores above .8 considered as good agreement and above .67 as acceptable (Artstein

² According to the *Early New High German Dictionary* it was used in this sense by just a few authors during the 15th and 16th century. Other dictionaries – except for *Duden Universalwörterbuch* [General dictionary] (from 1996 onwards) and *Brockhaus-Wahrig Deutsches Wörterbuch* [German dictionary] (1982) – do not show this sense of *Maus*.

³ For illustration, assume that both experts are lazy and award their points completely at random. Both overall scores would very likely fall somewhere between 43% and 57%, at first glance suggesting fairly good agreement. Only a closer look at the individual criteria in the grid would reveal that they only agree on approximately half of all points – exactly the value that is to be expected due to chance.

& Poesio, 2008, p. 576). To apply kappa to our data, we interpret the evaluation of qualitative criteria as a combination of two binary decisions: one point is awarded for minimum requirements, another point if the criterion is fully met. In this way, we obtain a satisfactory overall kappa of .73, with individual scores of .67 to .80 for the entries from the writing experiment, and as high as .86 and .87 for the two *duden.de* entries. As expected, the overall scores obtained by the two experts are much more similar than the kappa values might suggest.

For the sake of clarity, Table 4 only shows the assessment of one of the two experts for the various entries from the writing experiments (1.1–4.2), two expert entries (5.1, 5.2) and a completely AI-generated entry with expert prompting (6.1). In addition to the overall score, the table also shows how many points have been awarded in each general area and for individual categories. The colour coding is based on the colours in Table 3 and rates each category with one of the Grades I–V.

Table 4: Assessment of all dictionary entries according to the evaluation grid.

Source		writing experiments										Duden		AI
Entry		1.1	1.2	1.3	2.1	2.2	2.3	3.1	3.2	4.1	4.2	5.1	5.2	6.1
Using AI		y	Y	y	y	y	y	n	n	n	n	–	–	y
Lex. introduction		y	Y	y	n	n	n	y	y	n	n	–	–	–
Made by experts		n	N	n	n	n	n	n	n	n	n	y	y	y
%	100	46	58	57	40	50	40	59	46	32	43	70	78	90
Points	136	63	79	77	55	68	55	80	62	43	59	95	106	122
Whole microstructure	18	5	17	15	13	16	10	16	13	12	15	18	18	18
Textual structure / design	8	2	8	8	8	8	5	8	6	6	8	8	8	8
Overall rating	10	3	9	7	5	8	5	8	7	6	7	10	10	10
Comment on form	32	15	16	17	9	15	14	14	11	14	8	32	32	32
Lemma	3	3	3	3	3	3	3	3	3	3	3	3	3	3
Pronunciation	6	1	0	3	3	2	3	0	2	0	0	6	6	6
Grammar	14	11	10	11	3	10	8	11	6	11	5	14	14	14
Orthography	6	0	0	0	0	0	0	0	0	0	0	6	6	6
Frequency	3	0	3	0	0	0	0	0	0	0	0	3	3	3
Comment on semantics	86	43	46	45	33	37	31	50	38	17	36	45	56	72
Polysemy	4	0	2	3	2	2	2	4	4	4	4	3	3	4
Pragmatics	6	0	0	5	0	0	0	0	0	0	5	6	5	6
Meaning	16	14	11	11	9	11	12	12	12	13	13	11	14	13
Examples	14	11	10	11	9	9	11	14	11	0	14	10	13	12
Collocations	14	0	9	0	0	0	0	0	0	0	0	0	0	12
Word formation	10	8	0	0	0	0	0	10	6	0	0	0	0	8
Proverbs & idioms	5	0	0	0	0	0	0	0	5	0	0	5	5	4
Synonyms	10	10	9	9	7	9	0	10	0	0	0	4	10	7
Etymology	7	0	5	6	6	6	6	0	0	0	0	6	6	6

A first, striking observation is that only four of the ten dictionary entries from the writing experiment meet the minimum quality requirements according to Table 3. With one exception, they are from subjects that have received a lexicographical introduction. Nevertheless, even these entries remain in the lower range of Grade IV.

Due to space constraints, we can only discuss the best AI-generated dictionary entry 1.2 from the writing experiment in detail. Comment on form: Neither spelling nor pronunciation are provided. While all expected grammatical information is included, the inflectional paradigm is incomplete. The duplicated specification of the plural form is redundant. Content, orthography, and grammar are correct. Comment on semantics: Categories word formation, pragmatics and proverbs & idioms are missing. The meaning indication only covers three of the four required (sub)-senses, whereas sense 3 in the entry is questionable: “scheuer Mensch” [shy person]. Meaning indications for the main senses *rodent* and *computer input device* are very short or completely absent. Semantic categorisation is present at least for sense 1, but uses technical terms: “Familie Muridae” [family Muridae]. Each sense includes both an example and several collocations. Those of senses 1 and 2 appear authentic and appropriate; one of the collocations for sense 3 (“Angst vor Mäusen haben” [being afraid of mice]) obviously refers to another sense. In general, some collocations have doubtful empirical support due to lack of salience, e.g., “Mäuse loswerden” [getting rid of mice]. Sources for the examples are never provided in AI-generated entries. Synonym information is available for senses 1 and 2. Etymology information is also provided but has low information density. Except for the aspects mentioned above, all information is correct in terms of content and language. Regarding general criteria, the entry contains microstructure indicators and other structural elements that give the impression of a dictionary entry. The entry is written clearly, fulfils the dictionary functions and takes target users into account.

5.2 Comparison With Other Dictionary Entries

All entries from the writing experiments achieve much lower scores than the expert dictionary entry from *duden.de* (5.1 in Table 4). This was to be expected, as a short lexicographical introduction cannot make up for the decades of experience of professional lexicographers. Table 4 highlights some interesting details, e.g., for the comparison with dictionary entry 1.2: Whereas the expert entry achieves full marks in the areas of whole microstructure and comment on form, entry 1.2 falls far short of the requirements, as various categories are missing or incomplete. Both entries have almost the same score for the comment on semantics (46 vs. 45 points), mainly due to the missing information mentioned above. Information they do provide (meaning, examples) is similar in quality. A closer look at the qualitative differences between expert and AI-assisted dictionary entries is beyond the scope of the present paper.

5.3 Analysis of the Writing Experiment

In our writing experiment, the test subjects also completed a questionnaire with questions about their prior experience with lexicography/dictionaries, AI, and the

writing task. With two exceptions, all subjects had only moderate to little experience with both dictionaries and AI. Interestingly, entry 1.2 was generated by a subject who reported extensive experience with lexicography, while entry 2.1 from a subject with a high level of AI experience lags far behind at 40%. This suggests that AI experience may have little positive effect on the quality of an entry. There is no room here for a more detailed analysis of the questionnaires; establishing possible correlations between the quality of the generated texts and the level of knowledge of the subjects will require an experiment at a much larger scale. A comparison of the prompts used by the different groups is also beyond the scope of the paper.

5.4 AI-specific Considerations

In this Section, we discuss questions concerning the limitations of AI-assisted dictionary writing and the specific errors made by ChatGPT (1). We also address the question of prompting and, closely related, compare our findings to a purely AI-generated dictionary entry prompted by an expert (2).

(1) In general, GPT-3.5 is unable to generate audio snippets, as well as images or word clouds – features that are found in modern online dictionaries. Within the scope of our experiment, the limitations of AI are apparent mainly in the lack of comprehensiveness and the unclear structuring of the generated entries. Both issues can be solved by adequate prompting (see below). Comprehensiveness refers not only to the number of information categories, but also to the data within a category, e.g., missing forms of an inflectional paradigm, or missing word senses in the polysemy indication. Concerning the latter, ChatGPT often generates unclear, non-existent, rare or highly specialised senses such as “Scheue Person” [shy person] (entry 1.1) or “Biologisches Modell” [biological model] (entry 2.2). This can also happen for entries written by experts, though, as shown in Section 4.3. Similarly, meaning indications often use specialised technical terms, which is inappropriate for a general dictionary. This did not happen in the expert-generated entry, although it was not specifically forbidden. Further problems include giving not enough or too much information. The group without lexicographical introduction in particular created entries with redundant and unnecessary information, which contributed to disqualifying them as dictionary entries.

(2) The complex prompt below was developed by an expert to generate the optimized entry 6.1 in Table 4, which achieved an impressive overall score of 90% in the assessment:⁴

Generate a monolingual dictionary entry in German for the word Maus [mouse]; including the following categories: Lemma [lemma], Aussprache [pronunciation], Betonung [accentuation], Wortart [part of speech], Genus [genus], Deklination [declination], Rechtschreibung [spelling], Worttrennung [hyphenation], Häufigkeit [frequency], Bedeutung [meaning] (4) and just in brackets (pragmatische Einordnung) [(pragmatic classification)], Beispiele

⁴ One has to keep in mind, though, that the training data of GPT-3.5 very likely include several entries on the lemma *Maus* [mouse] from German online dictionaries.

[examples], Redewendungen und Sprichwörter [idioms and proverbs], Kollokationen [collocations] (sorted by meanings), Synonyme [synonyms], Wortbildung und Etymologie [word formation and etymology]. Please highlight the category names using bold type.

A closer look at Table 4 shows that the entry achieves full marks in the first two areas. It loses 14 points in the comment on semantics, which is due *inter alia* to the missing sources for examples, inaccuracies in the collocational information (“eine Mausfalle aufstellen” [set up a mousetrap] is a collocation of the compound *Mausefalle* [mousetrap]) and atypical synonyms. Nevertheless, it is evident that AI can in fact generate a highly-rated dictionary entry, which even surpasses an expert entry in term of completeness and breadth of the lexicographic information, if it is provided with a suitable optimised prompt. Consider the prompt for entry 1.2 by comparison, which obtained a much lower score:

ich muss Wörterbuchartikel zu den Stichwörtern das Nomen Maus und das verb köpfen verfassen. kannst du das für mich so machen, wie die Wörterbuchartikel in duden online aussehen. also zuerst information über Grammatik, z.B. Genus, Pluralbildung, Deklination usw. und dann Bedeutung mit jeweils 5 beispiele und auch kollokationen. Es wäre super, wenn du auch verwandte wörter darunter synonyme, antonyme, und andere verwandte wörter angibst. Die Wörterbuchartikel müssen auch mit Angaben über Frequenz und Herkunft versehen werden. Verfasse mir die Artikel für Maus und köpfen.

kannst du die informationen über die grammatik verbessern. nenne die überschriften so dass alle menschen die verstehen können. gib auch informationen darüber wie man genitiv bilden kann. kannst du versuchen die angaben zur herkunft für bestimmte bedeutung zu nennen und nicht für lemma⁵

The biggest differences between the two prompts are (a) language of the prompt, (b) structure of the prompt, (c) linguistic style and self-corrections. Whereas the expert combines English (instructions) and German (content) in his prompt, the test subject only uses German. His style of dialogue anthropomorphises the AI, addressing it directly (“kannst du” [can you]) rather than giving commands (“Generate”). His prompt is written in conceptually spoken form and can be described as linear-associative. In terms of content, the prototypical structure of a dictionary entry is not provided to the AI. The second prompt continues the dialogue rather than reformulating the original prompt. Experience from optimising the expert AI entry shows that the extension of a functioning prompt gives much better results than asking the AI to rewrite and improve its first draft, contrary to popular recommendations.

⁵ I have to write dictionary entries on the noun Maus [mouse] and the verb köpfen [behead]. can you do it for me like the dictionary entries in duden online. so first information about grammar, e.g., genus, pluralisation, declension etc. and then meaning with 5 examples each and also collocations. It would be great if you could also include related words, including synonyms, antonyms and other related words. The dictionary entries must also include information on frequency and origin. Write the entries for Maus and köpfen for me. | can you improve the information about the grammar. name the headings so that all people can understand them. also give information about how to form genitives. can you try to give information about the origin for specific meanings and not for the lemma [our translation]

5.5 Conclusions

For a final synthesis of the analysis, three main conclusions can be drawn (exemplified here for entry 1.2): (1) Expertise in relevant types of specialised texts often leads to better results, as the comparison with entries 2.1–2.3/4.1/4.2 shows. (2) Concise and targeted prompts, which are expanded in the correction phase, also produce better results, as the comparison with expert-prompted AI (6.1) shows. (3) Compared to expert dictionary entries (5.1), structural and qualitative differences are particularly noticeable. Based on these conclusions, we can give some specific recommendations for the use of AI chatbots in generating dictionary entries: (a) Prompting: We suggest to use complex prompts in the style of a command, containing all needed information in a structured manner, accompanied by specific instructions pertaining to the desired linguistic style and design. In the event that the results are deemed unsatisfactory, the entire prompt should be adjusted, reformulated, and used to generate a new version of the entry (rather than asking ChatGPT to revise the entry within the same thread); otherwise, the AI tends to retain the format of the last output and just adds or changes single pieces of information. (b) Output: Regardless of how prompts are crafted, we strongly recommend not to trust the generated data uncritically. A verification of the output against existing printed or online dictionaries is essential to assess the correctness of the generated content. In an educational context, it has the additional benefit of promoting dictionary literacy while demonstrating to users how actual dictionaries produced by experts function. (c) Our experiments indicate that using AI chatbots to create a dictionary entry can only be successful if the human user has a high level of lexicographic knowledge and text competence. Only against this background, a well-founded evaluation of the output can take place, which enables the user to make necessary revisions and improve prompts.

6. Outlook

Generative AI has reached a point where it takes more than just a brief, superficial glance at texts to recognise that they may not have been written by a human. A detailed (but time-consuming) qualitative evaluation that concentrates on the content and semantic aspects of the text, in terms of categories specific to the text type, provides much better insights, and is to be preferred to the superficial quantitative evaluation common in AI research. The elaborate evaluation grid proposed in this paper is a first step towards putting such thorough evaluation into practice.

Methodologically, our multi-level analysis framework still poses many challenges and unanswered questions. It is already clear, though, that it enables a relatively comprehensive evaluation of dictionary microstructure if the intended functions of the dictionary and its target users are taken into account. One of its key advantages is that the different evaluation categories and individual criteria are listed and assessed separately. This makes for a much more detailed and informative comparison of dictionary entries than mere overall evaluation scores. Of course, the evaluation grid needs to be dynamically adapted to the respective dictionary type and function. We have already done this for bilingual Chinese-German dictionary entries (Rink et al.,

2024), which enabled us to identify different types of errors made by the AI in a highly systematic manner. This paves the way to a better use of AI-generated dictionary entries in language teaching and learning.

We believe that our multi-level analysis framework and evaluation grid can serve as a general model for assessing and improving AI-generated texts of all types.

References

- Abel, A., Glaznieks, A., Linthe, M., & Wolfer, S. (Eds.) (2020). Textqualität im digitalen Zeitalter. *Themenheft Deutsche Sprache*, 48, 97–100. <https://doi.org/10.37307/j.1868-775X.2020.02>
- Arias-Arias, I., Domínguez Vázquez, M. J., & Riveiro, C. V. (2024). Efficiency and Intelligence in Lexicography and Artificial Intelligence: Can ChatGPT Recreate the Lexicographical Text Type?. *Lexikos*, 34(1), 51–76. <https://doi.org/10.5788/34-1-1879>
- Artstein, R., & Poesio, M. (2008). Survey entry: Inter-coder agreement for computational linguistics. *Computational Linguistics* 34(4), 555–596.
- Becker-Mrotzek, M. (2014). Schreibleistungen bewerten und beurteilen. In H. Feilke, & T. Pohl (Eds.), *Deutschunterricht in Theorie und Praxis. (DTP). Handbuch zur Didaktik der deutschen Sprache und Literatur in elf Bänden. 4. Schriftlicher Sprachgebrauch – Texte verfassen* (pp. 501–513). Baltmannsweiler.
- Celikyilmaz, A., Clark, E., & Gao, J. (2020). Evaluation of text generation: A survey. arXiv:2006.14799. <https://doi.org/10.48550/arXiv.2006.14799>
- de Schryver, G.-M. (2023). Generative AI and Lexicography: The Current State of the Art Using ChatGPT. *International Journal of Lexicography*, 36(4), 355–387. <https://doi.org/10.1093/ijl/ecad021>
- duden.de: *Duden – mehr als ein Wörterbuch*. Cornelsen Verlag GmbH. Retrieved May 31st, 2024, from <https://www.duden.de/>
- Engelberg, S., & Lemnitzer, L. (2009). *Lexikographie und Wörterbuchbenutzung*. Stauffenberg.
- Engelberg, S., & Storrer, A. (2016). Typologie von Internetwörterbüchern und -portalen. A. Klosa, C. Müller-Spitzer (Eds.): *Internetlexikografie. Ein Kompendium* (pp. 31–63). Walter de Gruyter.
- Fritz, G. (2017). *Dynamische Texttheorie*. Gießener Elektronische Bibliothek. <https://core.ac.uk/download/pdf/84117983.pdf>
- Lew, R. (2023). ChatGPT as a COBUILD lexicographer. *Humanities and Social Sciences Communications*, 10(1), 704. <https://doi.org/10.1057/s41599-023-02119-6>
- Mayring, P., & Hurst, A. (2017). Qualitative Inhaltsanalyse. In L. Mikos, & C. Wegener (Eds.), *Qualitative Medienforschung. Ein Handbuch* (pp. 494–502). UVK.
- Neumann, A. (2017). Zugänge zur Bestimmung von Textqualität. In M. Becker-Mrotzek, J.

Grabowski, & T. Steinhoff (Eds.), *Forschungshandbuch empirische Schreibdidaktik* (pp. 203–219). Waxmann.

Nussbaumer, M. (1991). *Was Texte sind und wie sie sein sollen. Ansätze zu einer sprachwissenschaftlichen Begründung eines Kriterienrasters zur Beurteilung von schriftlichen Schülertexten*. Max Niemeyer.

Nielsen, S. (2009). Reviewing printed and electronic dictionaries. A theoretical and practical framework. In S. Nielse, & S. Tarp (Eds.), *Lexicography in the 21st Century. In honour of Henning Bergenholtz*. (pp. 23–41). John Benjamins.

Pearsons, E., & Nichols, W. (2013). Toward a Framework for Reviewing Online English Dictionaries. In *Journal of the Dictionary Society of North America*, 34(1), pp. 201–210.

Ripfel, M. (1989). *Wörterbuchkritik. Eine empirische Analyse von Wörterbuchrezensionen*. Max Niemeyer.

Rink, C., Ganslmayer, C., & Evert, S. (2024) Towards a comprehensive method for evaluating and utilizing AI-generated bilingual lexicographical data in language learning using the example of Chinese as a foreign language. (ASIALEX. 2024. Proceedings), in print.

Sieber, P., & Nussbaumer, M. (1994). Texte analysieren mit dem Zürcher Textanalyseraster. In Sieber, P. (Ed.), *Sprachfähigkeiten – Besser als ihr Ruf und nötiger denn je!* (pp. 141–186). Sauerländer.

Svensén, B. (2009). *A Handbook of Lexicography. The Theory and Practice of Dictionary-Making*. Cambridge University Press.

Tarp, S. (2017). Dictionary criticism and lexicographical function theory. In M. Bielińska, & S. J. Schierholz (Eds.), *Wörterbuchkritik – Dictionary Criticism*. (pp. 113–132). Walter de Gruyter.

Wiegand, H. E. (1989). Arten von Mikrostrukturen im allgemeinen einsprachigen Wörterbuch. In H. E. Wiegand, & H. Steger (Eds.), *Dictionaries. An International Encyclopedia of Lexicography: Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography* (pp. 556–568). Walter de Gruyter.

Wiegand, H. E. (1998). *Wörterbuchforschung. Untersuchungen zur Wörterbuchbenutzung, zur Theorie, Geschichte, Kritik und Automatisierung der Lexikographie*. Walter de Gruyter.

Wiegand, H. E. (2010). Systematische Einführung. In H.E. Wiegand, M. Beißwenger, R.H. Gouws, M. Kammerer, A. Storrer, & W. Wolski. (Eds.), *Wörterbuch zur Lexikographie und Wörterbuchforschung. Dictionary of Lexicography and Dictionary Research*, Bd. 1 (pp. 1–121). Walter de Gruyter.

Contact information

Stephanie Evert

Friedrich-Alexander-Universität Erlangen-Nürnberg
stephanie.evert@fau.de

Christine Ganslmayer

Friedrich-Alexander-Universität Erlangen-Nürnberg
christine.ganslmayer@fau.de

Christian Rink

Friedrich-Alexander-Universität Erlangen-Nürnberg
christian.rink@fau.de

Carolina Flinz, Daniel Henkel, Valeria Zotti, and
Sabrina Ballestracci

A MULTILINGUAL PARALLEL CORPUS FOR THE LEXICAL INFORMATION SYSTEM LBC Recent Progress and Future Perspectives

Abstract The LBC-Platform (<https://www.lessicobeniculturali.net>) is a comprehensive lexical information system that aims to integrate various types of corpora and resources: dictionaries, concordances, monolingual Language for Special Purposes (LSP) corpora in different languages and LSP parallel corpora. Designed for users interested in cultural heritage, the platform provides free access to resources and tools such as NoSketch Engine, facilitating an in-depth understanding of the language of Fine Arts — encompassing painting, architecture, and sculpture — and highlighting its multidisciplinary nature and diverse discourse types. Numerous studies have demonstrated the platform’s utility for various purposes, including lexicography, translation, and didactics. This paper focuses on the preparation of the multilingual Vasari parallel corpus, part of the LBC Project’s four planned subcorpora (Vasari corpus, Literature corpus, Museum Web-Site Corpus, Travel Guide Corpus). Special attention is given to the Italian-English, Italian-French, and Italian-German components of the corpus, based on G. Vasari’s *Le vite de’ più eccellenti architetti, pittori, et scultori italiani* (1568) and its translations. In addition to methodological choices including alignment strategies and potential technical solutions, a small case study on the words *disegno* (‘drawing’) and *disegnare* (‘to draw’) will serve to illustrate the current status of the project and its relevance for LBC dictionaries, as well as future perspectives.

Keywords cultural heritage; parallel corpora; Language for Specific Purposes; artistic lexicon; translation

1. Introduction¹

Corpora of different kinds (LSP monolingual comparable corpora, LSP parallel corpora) form the core of the LBC-Platform,² a lexical information system (Klosa, 2016) created for different types of users interested in cultural heritage (Billero, Farina, & Nicolás Martínez, 2020) by the “Multilingual Cultural Heritage Lexicon” Research Unit of the University of Florence, in collaboration with scholars from Italian and foreign universities (Bologna, Paris 8, Pisa, Milan etc.). Resources (corpora, concordances)

¹ The four authors have written the paper jointly. Carolina Flinz and Sabrina Ballestracci are responsible for §1 and §2.3.3, Daniel Henkel for §2.2 and §2.3.1, Valeria Zotti for §2.1 and §2.3.2. Sections §3 and §4 were written jointly. In particular Carolina Flinz and Sabrina Ballestracci are responsible for the Italian-German part, Daniel Henkel for Italian-English and Valeria Zotti for Italian-French. We would also like to thank the trainees from the University of Florence (Sara Biagini, Carlotta Buricchi, Maria Chiara Susini) for the Italian-German alignment, as well as Valeria Zotti’s trainee from the University of Bologna, Elisa Squadrito, for the Italian-English alignment and part of the Italian-French one.

² Cfr. <https://www.lessicobeniculturali.net>

and tools (NoSketch Engine) are freely accessible for the user (Billero, 2020), providing a comprehensive vision of the language of Fine Arts (painting, architecture, and sculpture), characterised by multidisciplinary and interplay between discourse-types (Zotti, 2022).

The use and application of the comparable corpora present on the platform for different aims (lexicography, translation, didactics etc.) have been already demonstrated in different studies (most recently, see Ballestracci, Buffagni, & Flinz, 2024; Farina & Sini 2024; Natali, 2024).

A preliminary study by Loporini (2023), focusing on De Vere's English translation of Vasari's masterpiece, which had not been aligned yet at the time of the study (see 2.1), shows how information provided by general dictionaries can be enriched also by information from specialized parallel corpora. Analysing the word *disegno* ('drawing') and its collocations in Italian, together with a small number of translated segments³ into English, Loporini reveals the inherent polysemy of this lexeme, described as a "cline" from abstract (skill) to concrete (graphic representation), underlining the need for a parallel database to resolve translation difficulties.

The focus of this paper is on the preparation of the multilingual Vasari parallel corpus, one of the four subcorpora (Vasari corpus, Literature corpus, Museum Web-Site-Corpus, Travel Guide Corpus) of the LBC-Project which are in planning. In particular, attention will be given to the Italian-English, Italian-French and Italian-German components of the corpus based on a series of artists' biographies published in 1568, Giorgio Vasari's *Le vite de più eccellenti architetti, pittori, et scultori italiani* ("Lives of the Most Excellent Painters, Sculptors, and Architects," abridged in *Le Vite*) and its translations.

The discussion of methodological issues will encompass alignment (see 2.), possible technical solutions, the status of the project at present by giving a small case study (see 3.), its relevance for the LBC-dictionaries and future perspectives.

2. The Vasari Parallel Corpora: Methodology and First Results

2.1 The Source Text: *Le Vite* by Vasari

The decision to build a corpus on Vasari's *Le Vite* is due to its historic importance as the seminal text in Art History. In *Le Vite*, Vasari, a renowned Renaissance Italian painter and architect, describes the lives and works of Italian artists from the early 14th century (e.g., Cimabue, Giotto) to the second half of the 16th century (e.g., Leonardo da Vinci, Raphael and Michelangelo), for which he is recognized as the originator of Italian art criticism (Biffi, 2017).

Vasari published two editions of his work, the Torrentiniana in 1550, followed by the amplified Giuntina in 1568. In the 1568 edition, there are a total of 155 lives dedicated

³ Before sentence-level alignment had been completed, it was impossible to automatically extract source/target segment pairs, hence the use of a small sample of source/target pairs taken manually from both texts.

to various artists (about 834,000 words). Throughout a work as vast as *Le Vite*, which touches on many problems of a technical nature, such as artistic techniques, materials and tools, Vasari tried to conserve the technical vocabulary used by the artists who had preceded him (Le Mollé, 1988, p. 9). However, he did not have complete command of the Italian vocabulary, which was still evolving, and was sometimes obliged to ‘invent’ his own vocabulary to denote concepts that had either fallen into disuse, because the tools, techniques and recipes had been abandoned (Le Mollé, 1988, p. 7) or were entirely new and not yet stabilised in use. Vasari’s text is thus considered at the origin of the artistic lexicon.

His text has been translated into several languages over the centuries, thus contributing to the development of a pan-European lexicon of artistic terminology (Zotti & Henkel, 2024, p. 13). These translations are the means by which the Italian artistic proto-lexicon has been adapted to other languages. Translations are therefore a field in which specialists from different linguistic and cultural backgrounds explore the networks of meanings in Vasari’s text to forge new vocabulary and concepts (Dubus & Fiorato, 2017, p. 32). The analysis of translations of *Le Vite*, mostly based on the more complete Giuntina edition, reveals different lexical choices over almost five centuries in different languages.

The translations gave rise to a veritable ‘artistic terminology’ which has subsequently continued to evolve. Over the centuries, this specificity has fostered an abundance of cross-linguistic synonyms, parasynonyms or hyponyms in translation which reflect the complexity inherent in describing an artistic heritage linked to a unique historical, social and geographical context. Whereas the Italian source text, however, has remained unchanged since the 16th century, the ongoing creation of a multilingual database of target texts belonging to different centuries is necessarily diachronic in essence. Most translations of *Le Vite* made in the 17th and 18th centuries focus only on the biographies of the artists known best in the culture of the target language. For example, one of the first translations in France in the 17th century, Daret (1651), concerns the life of the painter Raphael. However, given that abridged translations are much more difficult to align, requiring many more manual interventions to deal with untranslated portions, preference has been given to complete translations for the purpose of creating a parallel database.

2.2 Alignment Methodology

The constitution of the parallel corpus follows a common alignment methodology for all languages. Source and target texts are pre-aligned at sentence level using LF Aligner and manually proofread by student interns following a common protocol under the supervision of the project coordinators, with further quality control carried out using Okapi Checkmate to detect suspicious disparities in length between source and target segments, a potential sign of misalignment. After automatic (pre)alignment, the first step in the proofreading process focuses on empty cells, as these necessarily indicate non-correspondence between the source and target texts. Most often, correspondences can be found between sentences in the

source text and sentences or sentence fragments in the target text. However, when additions or omissions are found, these are marked respectively as #NoSource# or #NoTarget#. Subsequently, as the risk of misalignment is greater for groups of sentences than for single sentences, the second step in the proofreading process focuses on sentences which have been grouped together into a single segment. Although “segments” generally correspond to “sentences”, whenever sentences in the source text correspond to more than one sentence in the target text, LF Aligner groups them together automatically. When this occurs in the other direction, i.e., two or more sentences in the source correspond to a single sentence in the target text, a source segment containing several sentences would be incompatible with other translations in other languages or from other periods, as other translators could hardly be expected to have made exactly the same choice to combine more than one sentence in exactly the same places. In addition to correcting misaligned segments, the proofreading protocol thus involves re-establishing the original sentence divisions in the source text so that it can serve as a common denominator for other translations from different periods or into different languages, with modifications in sentence segmentation (groupings, partial sentences) restricted solely to the target text.

Following alignment and proofreading, both source and target texts are tagged for Part of Speech (POS) and lemma in TreeTagger and converted to several formats: UTF-8 text (.txt), spreadsheets (.ods and .xlsx) and Translation Memory eXchange (.tmx, a specific XML standard for encoding bilingual or multilingual documents) to ensure compatibility with the greatest possible number of software tools.

2.3 Current Status and First Results

In the following three sections we briefly outline the current status of the alignment work for the Italian-English corpus, the Italian-French corpus and the Italian-German corpus.

2.3.1 Italian-English

Although *Le Vite* has been translated into English at different periods (Aglionby, 1685; Foster, 1850 *inter alia*), De Vere’s 1912–1915 English translation in 10 volumes was selected for inclusion in the LBC parallel database as it is the only English translation currently available in electronic format and is also the closest to a full translation, with only one section omitted *Introduzione Di Messer Giorgio Vasari Pittore Aretino Alle Tre Arti Del Disegno* (33,693 words, 4.4% of the total).

The English translation by De Vere was fully aligned with the Italian source text in 2022–23. De Vere’s translation is largely literal, with very few additions or omissions, and therefore relatively easy to align. The main corrections involved sentence divisions, as very long sentences in Italian were often divided into shorter sentences in English, although occasionally English sentences overlapped with two sentences in Italian.

In the following passage, for example, while the last segment pair contains single sentences in both languages, the first source sentence in Italian (*E prima andarono in mala parte ... la memoria dell'une e dell'altre.*) corresponds in the English translation to a shorter sentence and the first half of a longer sentence (with tildes “~~~” marking the merger), the second half of which corresponds to the second sentence in Italian (Table 1).

Table 1: Example of alignment for Italian-English from the *Proemio delle Vite* (Vasari, 1568; De Vere, 1912–1915)

Italian (Vasari 1568)	English (Vasari 1912–1915, translated by De Vere)
(1) E prima andarono in mala parte la pittura e la scoltura, come arti che più per diletto che per altro servivano; e l'altra, cioè l'architettura, come necessaria e utile alla salute del corpo, andò continuando, ma non già nella sua perfezzione e bontà; e se non fusse stato che le sculture e le pitture rappresentavano inanzi agl'occhi di chi nasceva di mano in mano, coloro che n'erano stati onorati per dar loro perpetua vita, se ne sarebbe tosto spento la memoria dell'une e dell'altre.	And the first to fall into decay were painting and sculpture, as being arts that served more for pleasure than for use, while the other—namely, architecture—as being necessary and useful for bodily weal, continued to exist, but no longer in its perfection and excellence. ~~~ And if it had not been that the sculptures and pictures presented, to the eyes of those who were born from day to day, those who had been thereby honoured to the end that they might have eternal life, there would soon have been lost the memory of both;
(2) Là dove alcune ne conservarono per l'immagine e per l'inscrizioni poste nell'architetture private e nelle pubbliche, cioè negli anfiteatri, ne' teatri, nelle terme, negli acquedotti, ne' templi, negli obelisci, ne' colossi, nelle piramidi, negli archi, nelle conserve, e negli erarii, e, finalmente, nelle sepolture medesime; delle quali furono distrutte una gran parte da gente barbara et efferata, che altro non avevano d'uomo che l'effigie e 'l nome.	whereas some of them survived in the images and in the inscriptions placed in private houses, as well as in public buildings, namely, in the amphitheatres, the theatres, the baths, the aqueducts, the temples, the obelisks, the colossi, the pyramids, the arches, the reservoirs, the public treasuries, and finally, in the very tombs, whereof a great part was destroyed by a barbarous and savage race who had nothing in them of man but the shape and the name.
(3) Questi fra gli altri furono i Visigoti, i quali avendo creato Alarico loro re, assalirono l'Italia e Roma, e la saccheggiorno due volte e senza rispetto di cosa alcuna.	These, among others, were the Visigoths, who, having created Alaric their King, assailed Italy and Rome and sacked the city twice without respect for anything whatsoever.

In each case, the necessary changes in sentence division have been restricted to the target text only, so that every segment pair corresponds to exactly one complete sentence in the original Italian source text.

2.3.2 Italian-French

For the alignment of the Italian-French database a first test (Zotti, 2017) was carried out manually for four French translations of Vasari's *Vita di Michelagnolo Buonarroti*, the longest biographical account and the most often studied of the different biographies. Given Michelangelo's well-known versatility, the analysis of this section proved a valuable source of lexical and artistic knowledge in the fields of painting, sculpture and architecture. This test showed the value of building a diachronic parallel corpus for qualitative analysis of artistic terminology as it has evolved over time (Zotti & Henkel, 2024).

Since 2021, the preparation of a parallel corpus including two different French translations (Weiss, 1903; Leclanché & Jeanron, 1839–1842) has been under way, but

is proceeding very slowly as the alignment process has encountered problems due to the nature of the translations: in one case the translation is abridged, while the other translation is rather free in terms of sentence structure.

The first French translation to be selected for alignment was that of Charles Weiss (1903), judged by the French specialists of the team to be the best in terms of quality (Farina & Billero, 2020). This translation, which was not available in digital format, was first converted into electronic format from the OCR-generated file available on the Internet Archive and then corrected manually by the LBC-Vasari team on Wikisource, where it is now available in digital format (Weiss, 1903). Despite its superior stylistic quality compared with other French translations, Charles Weiss' translation is incomplete: one biography is missing from the first part, five from the second and thirty-nine from the third (approx. 30% of the full text). For this reason, the decision was made to first supplement the biographies missing in Weiss (1903) with those taken from the complete translation by Leclanché & Jeanron (1839–1842), and then to complete the second alignment, once the proofreading of Leclanché & Jeanron (1839–1842) was finished.

The alignment process showed that Weiss' text contains many passages best described as “summaries” which proved exceedingly difficult to align as clear correspondences could not be established with specific sentences in the source text (cf. Table 2). The following passage shows several problems commonly encountered in this translation:

Table 2: Example of alignment for Italian-French from the *Vita di Giotto* (Vasari, 1568; Weiss, 1903)

Italian (Vasari, 1568)	French (Vasari, 1903, translated by Weiss)
(1) Questa opera dagl'artefici pittori era molto lodata, e particolarmente da Michelagnolo Buonarroti, il quale affermava, come si disse altra volta, la proprietà di questa istoria dipinta non potere essere più simile al vero di quello ch'ell'era.	Cette œuvre qui était très admirée par tous les peintres, particulièrement par Michel-Ange Buonarroti,
(2) Questa tavoletta, dico, essendo venuta in considerazione, da che si diede fuori la prima volta il libro di queste Vite, è stata poi levata via da chi che sia: che forse per amor dell'arte e per pietà, parendogli che fusse poco stimata, si è fatto, come disse il nostro poeta, spietato.	a été depuis enlevée par je ne sais qui, estimant peut-être qu'on ne l'appréciait pas assez[50].
(3) E veramente fu in que' tempi un miracolo, che Giotto avesse tanta vaghezza nel dipignere, considerando massimamente che egli imparò l'arte, in un certo modo, senza maestro.	#NoTarget#
(4) Dopo queste cose mise mano l'anno 1334 a dì 9 di luglio al campanile di S. Maria del Fiore; il fondamento del quale fu, essendo stato cavato venti braccia a dentro, una platea di pietre forti, in quella parte donde si era cavata acqua e ghiaia; sopra la quale platea, fatto poi un buon getto che venne alto dodici braccia dal primo fondamento, fece fare il rimanente, cioè l'altre otto braccia di muro a mano.	Le 9 juillet de l'an 1334, Giotto mit la main au campanile de Santa Maria del Fiore. ~~~ Après avoir fouillé le sol à la profondeur de vingt brasses, il établit comme fondation une couche de pierres dures, dans la partie qu'on avait asséchée, et y posa un massif de béton haut de douze brasses qu'il termina ensuite par huit brasses en pierres de taille.

In the first two segment pairs (1, 2), the corresponding segment in the target text, which is considerably shorter than the source sentence, is not a complete sentence syntactically and is incomplete semantically with respect to the information contained in the source sentence. The third source segment (3) has no counterpart whatsoever in the target text, while in the fourth segment pair (4), the source sentence corresponds to two sentences in the target text.

Such gaps and disparities occur continuously throughout Weiss' French text. Although Weiss' translation is more or less literal at the beginning of each biography, the omissions and reductions become more and more frequent as it progresses, resulting in serious discontinuities in the automatic alignment, which relies on segment length as one of the main criteria to identify corresponding segment pairs.

Because of these problems, the second translation by Leclanché & Jeanron (1839–1842), which is the first complete French translation of Vasari's *Le Vite*, is currently undergoing proofreading and alignment. This translation is analytical, based on the tradition of earlier studies of the work, since the two translators took into account the translations that already existed in England and Germany, with many notes alongside the French version of Vasari's original text. The French translation, *stricto sensu*, mainly by Léopold Leclanché, a lawyer and politician, is in fact enriched by a vast array of commentaries by Jeanron, a painter, draughtsman and lithographer. These comments, placed at the end of each Artist's biography, are marked with a paragraph separator symbol which will help to detach them from the text itself, thus facilitating alignment of the full version. Leclanché & Jeanron's translation, which corresponds almost perfectly to the Giuntina edition of 1568, even though Leclanché often rephrases to explain certain artistic concepts for which no specific terms existed in Vasari's time, laid the foundations for future translations of Vasari's text (Zotti, 2017), all of which are based on this edition, which they improve especially with respect to terminological accuracy (Weiss, 1903; and later Chastel, 1981).

2.3.3 Italian-German

For German, Kupper's 2008 edition was selected. It is based on the first German translation of *Le Vite* that has come down to us, edited by Schorn and Förster in 1832–1849, to which Kupper made some minor changes, such as normalising spelling according to new conventions. Unlike many earlier editions, this is an almost complete translation, despite certain omissions as revealed in the alignment process.

The first attempt at alignment highlighted some shortcomings of the text in German translation. In particular, it came to light that a few chapters of the initial part of the Italian text had not been translated, namely: the dedication *Agli artefici del disegno*, the *Proemio di tutta l'opera* and the *Introduzione di messer Giorgio Vasari pittore aretino alle tre arti del disegno, cioè architettura pittura e scoltura e prima dell'architettura*. As these parts were at the beginning, a new alignment phase was necessary, conducted by removing the parts not included in the German text from the Italian text and then inserting them again after the new automatic (pre)alignment.

The analysis of the alignment revealed a small number of #NoSource# cells (additions), but a larger number of #NoTarget# cells (omissions). The former are principally cells that correspond to quotations in the source text, which in the target text are also translated into German. The latter are mainly anecdotes about the lives of artists, which are omitted or summarised, i.e., condensed into fewer sentences. Moreover, where there are complex sentences in Italian, the German translation tends to divide them into several sentences of greater simplicity:

Table 3: Example of alignment for Italian-German of the *Vita di Cimabue* (Vasari, 1568; Kupper, 2008)

Italian (Vasari, 1568)	German (Vasari, 2008, translated by Kupper)
(1) Avendo poi preso a fare per i monaci di Vall’Ombrosa nella Badia di Santa Trinita di Fiorenza una gran tavola, mostrò in quell’opera, usandovi gran diligenza per rispondere alla fama che già era conceputa di lui, migliore invenzione, e bel modo nell’attitudini d’una Nostra Donna, che fece col Figliuolo in braccio e con molti Angeli intorno che l’adoravano in campo d’oro; la qual tavola finita, fu posta, da que’ monaci in sull’altar maggiore di detta chiesa, donde essendo poi levata, per dar quel luogo alla tavola che v’è oggi di Alesso Baldovinetti, fu posta in una cappella minor della navata sinistra di detta chiesa.	Darauf übernahm er für die Mönche von Vallombrosa in der Abtei von Santa Trinita in Florenz eine große Tafel. Er wendete großen Fleiß auf dies Werk, um dem Ruf zu genügen, den er sich schon erworben hatte, und zeigte darin noch viel bessere Erfindung und schöne Stellungen. ~~~ Es war eine Mutter Gottes, mit dem Kind auf dem Arm und vielen Engeln umher, welche sie anbeten, auf Goldgrund. ~~~ Dieses Bild stellten jene Mönche über dem Hauptaltar ihrer Kirche auf, von wo es später weggenommen und in eine kleinere Kapelle des linken Seitenschiffs der Kirche gebracht ward, um dem Gemälde von Alessio Baldovinetti Platz zu machen, welches noch heute auf jenem Hauptaltar steht.

In the future, the corpus will be expanded to include the translation by the *Kunsthistorisches Institut* (KHI) in Florence (2004–2015), a translation that stands out from all previous ones because of its focus on Vasari’s style and specialised vocabulary.

3. Case Study: *disegno* / *disegnare* and Their Translation Counterparts

As indicated above, Vasari’s parallel database will be integrated in the LBC-Platform whose one of the main objectives is facilitating the translation, in different foreign languages, of texts belonging to diverse discourse types in the domain of Fine Arts. The parallel corpus will thus certify new translation equivalents of the lemmas (words and collocations) which will be included in the future LBC-dictionary so as to provide quality lexicographical information (Farina & Billero, 2020), given that the multilingual lexicographic and lexical resources in print and online are not always adequate or helpful in identifying possible translation counterparts. In our small case study, we will concentrate on the lexemes *disegno* (‘drawing’, noun)⁴ and *disegnare* (‘to draw’, verb) in order to give a glimpse of this future innovative lexicographical application. These two lexemes have different variants of meaning, enriched over time, entering in the specialised lexicons of various disciplines or areas of experience (Ballestracci, 2023, p. 22). In Vasari’s masterpiece they have a complex and multifaceted semantic potential, characterised by multiple meanings. The frequency is high: the noun/

⁴ These lexemes have been studied previously by Luporini (2023) in one English translation and by Balestracci (2023) in the *Vita di Raffaello* and its German translation (Kupper, 2008 and KHI), using the LBC database while alignment of source/target texts was still in progress.

lemma *disegno* has 1433 occurrences and the verb/lemma *disegnare* 530 occurrences, among the most frequent nouns and verbs in the entire work (*disegno* is the 7th most common noun, following *cosa* ('thing'), *opera* ('work'), *figura* ('figure'), *mano* ('hand'), *tempo* ('time') and *anno* ('year'), while *disegnare* is the 30th most common verb).

Interesting patterns are revealed by examining their counterparts in the aligned corpora. In the English translation by De Vere (1912–1915), *disegno* corresponds in order of descending frequency to *design* (142), *drawing* (53), *draughtsmanship* (11), *draw* (8), *draughtsman* (3), *likeness* (1), *proof* (1), while *disegnare* is rendered with *to draw* (49), *to design* (16), *drawing* (10), *outline* (4), *draughtsman* (2), *draughtsmanship* (2), *to copy* (1), *to intend* (1), *to propose* (1), *to take* (1), *to trace* (1). These findings confirm Luporini's (2023) observations concerning the three most common interpretations of *disegno*: *design*, *drawing* and *draughtsmanship*, and offer two new (albeit much less common) alternative translations: *proof* and *likeness*. The same three most common counterparts of the noun *disegno* are present among the most common translations for the verb *disegnare*, as well as a fourth technical meaning corresponding to *outline*, *trace* and a fifth, non-technical interpretation rendered by *intend* or *propose*. It is interesting to note, with reference to Luporini's "cline" from abstract to concrete, that the verb *disegnare* is most often interpreted in the most concrete sense of *draw/drawing*, while the median interpretation *design* is more common for the noun *disegno*. For both, the most abstract interpretation corresponding to *draughtsmanship* is the least common of the three main meanings. The parallel database thus provides a comprehensive overview of the semantic profile of these words as revealed in their alternative translation counterparts.

In French, so far only the chapters known as the *Teoriche* (especially *Introduzione Di Messer Giorgio Vasari Pittore Aretino Alle Tre Arti Del Disegno*) have been semi-automatically aligned with Weiss (1903) for a total of about 50.000 words. In these chapters, which are missing from most other translations but are particularly interesting from a terminological point of view, the most common translation counterparts for *disegno* are: *dessin* (81), *dessiner* (3), *composer* (1), *composition* (1), *contour* (1), *dessein* (1), *modèle* (1), *propriété* (1). The verb *disegnare* most commonly corresponds to: *dessiner* (22), *dessin* (4), *chercher* (1), *marquer* (1), *reproduire* (1), *tracé* (1), *traiter* (1). The polysemy of *disegno* and *disegnare* is thus apparent in their translation counterparts.

In German, only a quarter of the Italian source text has been aligned so far (24,112 words). The lemma *disegno* (330 occurrences, of which 289 are of *disegno* and 50 of the plural form *disegni*) is translated mostly by *Zeichnung* ('drawing') (131) and *Zeichnungen* ('drawings') (32) often adding the corresponding word in Italian in square brackets [*disegno*], in those cases where the translated term has a more precise meaning than the German (see also Ballestracci, 2023, pp. 22–23). Other translation counterparts with substantives include *Zeichenkunst* (24) ('drawing art', mostly followed by the Italian *disegno* in square brackets in those cases); *Plan* ('plan'), when it is used with verbs like *entwerfen* ('to draft') or *zeichnen* ('to draw'); *Angabe/Angaben* (5) ('indication/indications'); *Blatt/Blätter* (3) ('sheet/sheets'); *Entwurf/Entwürfe* (2) ('draft/drafts'); *Modell* (1) ('model'); *Arbeit* (1) ('work') and *Werk* (1) ('work'). In some cases the lemma is translated with

verbs like *zeichnen* ('to draw') or substantive verbs like *Zeichnen* ('drawing'). In a few cases the lemma is not translated at all. Interesting cases, in which the lexeme has other meanings, are the translations of *disegno* with *nach richtigem Verhältnis* ('according to the correct ratio'), with *was hast du vor?* ('what are you planning to do?') and with *so gut* ('so good'). The verb *disegnare* (65) is translated mainly by the verb *zeichnen* (46) ('to draw') or with *ausführen* (3) ('to carry out'), *entwerfen* (2) ('to draft'), *malen* (1) ('to paint'), *darstellen* (1) ('to represent'). In some cases there the verb is translated with the nouns *Zeichnung* (3) ('drawing'), *Blatt* (1) ('sheet') with the phraseme *von seiner Hand* (1) ('by his Hand') or not translated at all. This quick overview of some of the most used translation counterparts in German highlights several potential meanings (more general, more specific, phraseological) of the lexemes in the source language.

4. Conclusion

Parallel corpora are not only important for the linguistic analysis of translation but also from a contrastive point of view: work on parallel corpora makes it possible to conduct quantitative and qualitative analyses on large quantities of data, focusing on Vasari's style and specialised vocabulary, comparing lexical and syntactic aspects between original texts and translations.

The polysemy of lexemes like *disegno* can be revealed in translation, delineating interlingual correspondences that are useful for specialist translation in the field of art. Careful qualitative analysis elucidates semantic peculiarities that are central for lexicographic purposes. In the specific case study presented herein, analysis of the translations shows that, over time, ambiguities in Vasari's use of 16th-c. Italian are replaced over time by greater terminological precision in other languages.

Indeed, the study of *disegno* and *disegnare* and their counterparts in English, French and German reveals that in Vasari, these lexemes were used with a broad range of meanings including, on the one hand, technical meanings close to terminology and, on the other hand, descriptive meanings belonging to the general language and also phraseological uses.

In English, De Vere tried to render these different meanings as precisely as possible, hence a greater lexical variety which can also be seen as a sign or consequence of the development of artistic terminology between 16th-c. Italian and 20th-c. English. Also, the French and German translators chose corresponding terms that have now become stabilized, according to their purpose and intended public.

Future work in the LBC project will involve the alignment of other translations. In French the alignment of Leclanché & Jeanron (1839–1842) will be completed, followed by the alignment of translation by a renowned French art historian, Chastel (1981). In German, for example, the alignment of the translation of KHI is planned, which contains even more counterparts for *disegno* highlighting its semantic complexity (Ballestracci, 2023, p. 30). The alignment of *Le Vite* with its translations in other languages of the LBC Project like Spanish and Russian is planned as well.

No comprehensive inquiry into translation in the field of the fine arts and art history has yet been undertaken, due to the technical difficulties inherent in creating diachronic parallel corpora, but the work carried out by the LBC team is an attempt to promote new research in this field which will have a clear impact also in the field of cultural heritage.

References

Aglionby, W. (1685). *Painting Illustrated in Three Dialogues. Includes the first English translation of Giorgio Vasari's writings on the lives of eminent artists*. John Gain.

Ballestracci, S. (2023). Da un segno a tanti segni. L'emergere della polisemia del termine disegno nelle traduzioni tedesche delle Vite di Vasari. In V. Zotti, & M. Turci (a cura di.), *Nuove strategie per la traduzione del lessico artistico* (pp. 21–38). Firenze University Press.

Ballestracci, S., Buffagni, C., & Flinz, C. (Eds.), (2024). Das deutsche LBC-Korpus: Zusammenstellung und Anwendung. In S. Ballestracci, C. Buffagni, & C. Flinz (Eds.), *Das deutsche LBC-Korpus*. Firenze University Press. <https://corpora.lessicobeniculturali.net/de/deutsche-LBC-korpus-zusammenstellung-anwendung>

Biffi, M. (2017). *L'italiano dell'arte, in L'Italiano. Conoscere una lingua formidabile*. Accademia della Crusca.

Billero, R., Farina, A., & Nicolás Martínez, M. C. (Eds.). (2020). *I Corpora LBC. Informatica Umanistica per il Lessico dei Beni Culturali*. Firenze University Press.

Billero, R. (2020). Implementazione di software per la gestione dei corpora LBC. In R. Billero, A. Farina, & M. C. Nicolás Martínez (Eds.), *I Corpora LBC. Informatica Umanistica per il Lessico dei Beni Culturali* (pp. 19–32). Firenze University Press.

Daret, P. (1651). *Abregé de la vie de Raphaël Sansio d'Urbain, tres-excellent peintre & architecte; ou il est traité de ses oeuvres, des stampes qui en ont esté gravées, tant par Marc-Antoine Bolognois qu'autres excellent [« sic »] graveurs*. Trad. en françois par P. Daret, graveur.

Dubus, P., & Fiorato, C. L. (2017). (Eds.). *La réception des Vite de Giorgio Vasari dans l'Europe des XVIe-XVIIIe siècle*. Librairie Droz.

Farina, A., & Billero, R. (2020). Corpus in “Natural” Language Versus “Translation” Language: LBC Corpora, A Tool for Bilingual Lexicographic Writing. In D. F. Iezzi, D. Mayaffre, & M. Misuraca (Eds.), *Text Analytics. JADT 2018. Studies in Classification, Data Analysis, and Knowledge Organization*. Springer, Cham. https://doi.org/10.1007/978-3-030-52680-1_14

Farina, A., & Sini, L. (2024). Le Corpus LBC Français. In A. Farina, *Corpus LBC Français*. Firenze University Press. <https://corpora.lessicobeniculturali.net/fr/corpus-LBC-francais>

Klosa, A. (2016). Der lexikographische Prozess im Projekt elexiko. In V. Hildenbrandt, & A. Klosa (Eds.), *Lexikographische Prozesse bei Internetwörterbüchern* (pp. 29–38). OPAL 1/2016 Institut für Deutsche Sprache.

Le Mollé, R. (1988). *Georges Vasari et le vocabulaire de la critique d'art dans les Vite*. Editions littéraires et linguistique de l'université de Grenoble.

LBC: *Lessico dei Beni Culturali*. Retrieved July 30, 2024, from <https://www.lessicobeniculturali.net>

LF Aligner. Retrieved July 30, 2024, from <https://sourceforge.net/projects/aligner/>

Luporini, A. (2023). Exploring the Lexis of Art Through a Specialized Corpus: A Bilingual Italian-English Perspective. *International Journal of English Linguistics*, 13(7), 28–38.

Natali, I. (2024). The English Corpus for the LBC project. In I. Natali, *The LBC English Corpus*. Firenze University Press. <https://corpora.lessicobeniculturali.net/en/english-corpus-LBC-project>

Okapi Checkmate. Retrieved July 30, 2024, from <https://okapiframework.org/wiki/index.php?title=CheckMate>

Vasari, G. (1550). *Le Vite de' più eccellenti Pittori, Scultori et Architettori*. Edizione Torrentiniana.

Vasari, G. (1568). *Le Vite de' più eccellenti Pittori, Scultori et Architettori*. Edizione Giuntina.

Vasari, G. (1832–1849). *Leben der ausgezeichnetsten Maler, Bildhauer und Baumeister. Von Cimabue bis zum Jahre 1567*, a cura di Ludwig Schorn e Ernst Förster. Cotta.

Vasari, G. (1839–1842). *Vies des peintres, sculpteurs et architectes*. Trad. et commentaires par Léopold Leclanché et Philippe-Auguste Jeanron. Tessier.

Vasari, G. (1850). *Lives of the most eminent painters, sculptors, and architects: translated from the Italian of Giorgio Vasari*. With notes and illus., chiefly selected from various commentators by Mrs. Jonathan Foster. Robarts – University of Toronto.

Vasari, G. (1903). *Les vies des plus excellents peintres, sculpteurs, et architectes de Giorgio Vasari*, traduction de Charles Weiss. Dorbon-Aine. Available on wikisource :<https://fr.wikisource.org/wiki/Les_vies_des_plus_excellents_peintres,_sculpteurs,_et_architectes>

Vasari, G. (1912–1915). *Lives of the Most Eminent Painters, Sculptors and Architects* (10 volumes; London: Macmillan and Co. and the Medici Society, 1912–1915), trans. by Gaston du C. De Vere.

Vasari, G. (1981). *La vie des meilleurs peintres, sculpteurs et architectes*, traduction d'André Chastel. Berger-Levrault.

Vasari, G. (2004–2015). *Deutsche Ausgabe der 'Vite' Giorgio Vasaris*, a cura di A. Nova con M. Burioni, K. Burzer, S. Feser, H. Gründler, & F. Jonietz. Wagenbach.

Vasari, G. (2008). *Leben der ausgezeichnetsten Maler, Bildhauer und Baumeister von Cimabue bis zum Jahre 1567*, a cura di D. Kupper. Directmedia Publishing.

Zotti, V. (2017). L'integrazione di corpora paralleli di traduzione alla descrizione lessicografica della lingua dell'arte: l'esempio delle traduzioni francesi delle Vite di Vasari. In V. Zotti, & A. Pano Alaman (Eds.), *Informatica umanistica. Risorse e strumenti per lo studio del lessico dei beni culturali* (pp. 105–134). Firenze University Press.

Zotti, V. (2022). La langue des Beaux-Arts: dimension interculturelle et enjeux terminologiques des corpus comparables. In C. Grimaldi et al. (Eds.), *Terminologia e interculturalità. Problematiche e prospettive* (pp. 229–245). I libri di Emil.

Zotti, V., & Henkel, D. (2024). The diachronic evolution of artistic terminology in translation. Building a parallel corpus of Giorgio Vasari's *Le Vite*. *CORPUS*, 25, 1–20.

Contact information

Carolina Flinz

Università degli Studi di Milano
carolina.flinz@unimi.it

Daniel Henkel

Université Paris 8
daniel.henkel@univ-paris8.fr

Valeria Zotti

Alma Mater Studiorum – di Bologna
valeria.zotti@unibo.it

Sabrina Ballestracci

Università degli Studi di Firenze
sabrina.ballestracci@unifi.it

Peter Juel Henriksen

MAKE EACH MORPH COUNT

A New Approach to Computational Lexicography for Text Processing

Abstract In this paper we present a newly developed formal framework (as well as its practical implementation) for automatic, lexically driven analysis of Danish text tokens. The framework (called “CLINK”) employs a minimal token definition (the “morph”) and a compact lexical representation (the “CLINK template”). All morphs (i.e., text elements with individual semantic contribution) are lexicalized using the same template, word forms, affixes, glue elements, punctuation marks, multi-word expressions, etc. Thus, the definition of “lexeme” is reinterpreted in functional-computational terms. The grammar rules of CLINK are purely abstract, viz. those of the Lambek calculus (categorical grammar). This paper gives an overview of the CLINK framework (motivations and application). References to performance metrics will be given (suggesting CLINK to be on a par with the Danish state-of-the-art in PoS-tagging while providing much richer annotation structure). However, we consider the formal framework in itself to be the main contribution of this short paper.

Keywords computational lexicography; language technology; text analysis; categorical grammar; CLINK

1. Introduction

Most dictionaries for human users have entries for words only, making them difficult to utilize in language technology, especially for applications aimed at text (such as grammar and spell checkers, machine translation and dialog systems). In this short paper, we propose a new approach to computational lexicography based on a generalized definition of lexeme, more specifically that any text element with an individual semantic contribution be lexicalized. We shall refer to such elements as lexical morphs (inspired by Haspelmath 2020). The class of morphs thus comprises lexical lemmas on a par with sub-verbal and non-verbal tokens (affixes, glue elements, punctuation marks, icons, and so forth). As a practical demonstration we present the text analyzer CLINK, reading a text and returning each token annotated with a link to a lexical entry (or possibly multiple links). As a lexical base we use the newly published Central Word Register for Danish (“COR”, cf. Henriksen, 2023; Dideriksen et al., 2023; Widmann 2024), complemented with dictionaries for non-verbal lexemes and more. As we shall argue, CLINK’ed text provides enhanced input for language technology.

2. The Central Word Register

Retskrivningsordbogen (“RO”, Jervelund et al., 2012) is the dictionary defining the Danish orthographic norm (Act 1997), available also as a machine readable database

called COR_1 for language technology. Each lexeme in COR_1 carries a unique id specifying its class (PoS or otherwise), as exemplified in Table 1. The database is available at <https://ordregister.dk>.

Table 1: COR_1 samples. Indices with 5 digits: lemma, 3 digits: PoS

Lexeme	Lemma	COR_1 -id	Inflection	Gloss
hus	HUS	COR.43962.120	neut.sg.indef	<i>house</i>
huset	HUS	COR.43962.121	neut.sg.def	<i>the-house</i>
huse	HUS	COR.43962.122	neut.pl.indef	<i>houses</i>
husene	HUS	COR.43962.123	neut.pl.def	<i>the-houses</i>
mus	MUS	COR.74798.110	com.sg.indef	<i>mouse</i>
mus	MUS	COR.74798.112	com.pl.indef	<i>mice</i>
cyber	CYBER	COR.02858.890	prefix	<i>cyber-</i>

Notice in Table 1 that “mus” is ambiguous in Danish between a singular and a plural reading; but the corresponding COR ids are not. In general, COR linking (i.e., annotating text tokens with COR ids) effectively disambiguates the homographs.

3. Morphological Analysis as Formal Deduction

CLINK’s formal framework is based on categorial grammar (CG). Historically, CG has mainly been used for syntactical analysis of the sentence and its constituents. CG logic is however quite general and just as fit for morphology, or indeed any complex analysis with a syntax, a semantics, and a compositional relation between the two (Wood, 1993, is a good CG primer for linguists; Morill, 2010, is more complete).¹

3.1 The Lexical Template

Each morph must be available to CLINK in the form of a lexical template. Formally, the template is a 5-tuple (**text-in**, **text-out**, **cat**, **sem**, **phon**). Observe in Table 2 that, for most morphs imported from COR_1 , the category (**cat**) is identical to the PoS index, and the semantic proxy (**sem**) to the lemma index; exceptions are function morphs (such as the prefix “cyber” and the grapheme “.”) not occurring as free forms. Categories containing a slash require an argument: $y \backslash x$ thus needs an argument y to its left (x/y to its right) in order to become x . For example, two adjacent categories 601 601\602 will reduce to 602.

¹ To the best of our knowledge, Lambek grammar has not been in use for commercial NLP since the early 1990es, and even then only in smallish and strictly rule based systems.

Table 2: CLINK templates. The pair **text-in/text-out** constitutes a difference list (as in Prolog or Haskell). C=any characters; P=punctuation marks; A=alphabetic characters; ⁰=possibly empty. X,Y=category. a,b=semantic term; s01,s02,... = semantic proxy for morphs not in COR₁; /.../ = phonetic forms (computer-readable symbols, Wells, 1997); crd,ord,pfunc = phonetic functions (notice that ‘silent’ morphs, even ‘’, may affect their surroundings).

lexeme	text-in	text-out	cat	sem	phon
‘mus’	[mus C ⁰]	C ⁰	110	43651	/mu:ʔs/
‘mus’	[mus P ⁰]	P ⁰	112	43651	/mu:ʔs/
‘hus’	[hus C ⁰]	C ⁰	120	43962	/hu:ʔs/
‘cyber’	[cyber A]	A	X/X	λa.a(02858)	/sAJbC/
‘e’	[e A]	A	X\Y/Y	λab.b(a)	pfunc(GLUEe)
‘’	A	A	X\Y/Y	λab.b(a)	pfunc(GLUE0)
‘6’	[6 P ⁰]	P ⁰	601	s06	/sEgs/
‘.’	[. P ⁰]	P ⁰	601\602	λa.ORD(a)	crd(a)>ord(a)
‘.’	[. P ⁰]	P ⁰	X\X*41	λa.[a,s41]	pfunc(fulstp)

3.2 Compound Analysis

Compound nouns like “cybermus”_{SG} (*cyber mouse*), “husmus”_{SG} (*house mouse*) and “musehus” (*house for mice*) are segmented into morphs as illustrated in Figure 1 (rows “SEG”).

SEG	[cyber]	[mus]		[cybermus]
SEQ	X/X	110	==>	110
SEM	λa.a(02858)	43651		43651(02858)
SEG	[hus]	[]	[mus]	[husmus]
SEQ	120	Q\Y/Y	110	==> 110
SEM	43962	λab.b(a)	43651	43651(43962)
SEG	[mus]	[e]	[hus]	[musehus]
SEQ	110	Z\W/W	120	==> 120
SEM	43651	λab.b(a)	43962	43962(43651)

Fig. 1: Compound analysis: segmentation, sequent, semantic form

Each segmentation is mapped to a sequent (rows “SEQ”) in the form ANTE==>CON, where ANTE (antecedent) specifies the list of categories, and CON (consequent) is the hypothesis to be proven. In case CON can be deduced from ANTE in formally valid steps,² the sequent has found its proof – and the

² The Lambek calculus has these seven proof rules:

(/L) $L0\ A/B\ M1\ N0 \implies C$ IF $M1 \implies B$ AND $L0\ A\ N0 \implies C$;

(\L) $L0\ M1\ B/A\ N0 \implies C$ IF $M1 \implies B$ AND $L0\ A\ N0 \implies C$;

(*L) $L0\ A*B\ M0 \implies C$ IF $L0\ A\ B\ M0 \implies C$;

(/R) $L1 \implies A/B$ IF $L1\ B \implies A$; (\R) $L1 \implies B/A$ IF $B\ L1 \implies A$;

(*R) $L1\ M1 \implies A*B$ IF $L1 \implies A$ AND $M1 \implies B$; (**axiom**) $A \implies A$;

A, B, C are categories, and L_n, M_n, N_n are lists of $n+$ categories.

Wood (1993) explains how to use the proof rules: For each rule *Seq* IF *Prem*, install a sequent for *Seq*, unify all variables, and you get one or more new premises *Prem*. Repeat for each premise until reaching **axiom**. You’re done.

compound word its category. In Figure 1, proofs are found for $X=110$, $Q=120$, $Y=110$, $Z=110$ and $W=120$. Observe that each proof is associated with a lambda-formula (rows “SEM”) specifying the semantic relations between the morphs (compare “husmus” and “musehus”).

3.3 Lexical and Structural Ambiguity

In general, homographs have distinct semantic projections.³ Danish “lyst” can, for example, be a verbal participle (*shined*), a noun (*desire*) or an adjective (*bright*), all three pronounced differently. Approximately 14.5% of COR₁’s word forms are homographs, many of these highly frequent. Among the ten most frequent tokens in typical text,⁴ eight are homographs in COR₁ (“i”, “en”, “til”, “af”, “på”, “at”, “det”, “for”), token “for” with no less than seven entries.

Danish is, like German, a compounding language. Words like “husmusehus” are readily created, in this case ambiguous between the meaning *house for domestic mice* or *mouse house for in-house use*. The reader may wish to confirm, CLINK-style, the two projections 43962(43651(43962)) and (43962(43651))(43962).

The Danish rules of interpunction are another source of ambiguity. The grapheme “.” is lexically ambiguous between a full stop with sentential scope (as in “Klokken er 6.”, *It’s 6 o’clock.*) and an ordinal suffix with local scope (as in “på 6. sal”, *on the 6th floor*). In the latter case, the morphs “6” and “.” do combine morphologically; but not in the former. The projections are [s06,s41] and ORD(s06), respectively, the former simply listing the arguments, the latter composing them.

4. Computers Prefer Lambdas

Computers like COR-ids and λ -forms far better than text tokens. This is a challenge for any NLP project. Even state-of-the-art MT systems (machine translation) and chatbots often choke on lexical and structural ambiguities. Consider e.g., the Danish sentence “husmus bor i hus”, translating naturally into *house mice live in houses* or even *domestic mice are living indoor*. Google Translate, in contrast, suggests “house mouse lives in house” – neither accurate nor pretty. In the same vein, Danish synthetic voices are notorious for mispronouncing homographs like “for”, “sig”, “lyst”, “så”. Last but not least, Danish spell and grammar checkers (MS-Word, Google Docs, LibreOffice) score poorly on compound recognition and punctuation. We believe that preprocessing the input text CLINK-style could solve most of these issues. Of course, this requires a CLINK parser comparable in performance to the state-of-the-art for Danish language taggers. However, CLINK being an open standard, we can hope for a wider scope of contributions. More on this below.

³ Notice that “mus”_{SG} and “mus”_{PL} both project to 43651 as they belong to the same lemma and hence share the same proxy. In fig.1, “mus”₁₁₂ would also produce a proof (for $Z=112$), however with the same semantic projection as for “mus”₁₁₀

⁴ All texts in DaGW (Danish Gigaword Corpus, see Derczynski et al. 2021).

4.1 Text Annotation

CLINK (standard configuration) annotates each input token for **cat** (proven consequent) and **sem** (λ -form). Consider an example from the classical Danish prose poem *Ved Frokosten* (fig. 2).

Gud	[gud]	900:0	14223:0
ske	[ske]	900:1	14223:1
Lov	[lov]	900:2	14223:2
for	[for]	880	00093
Sofahjørnets	[sofa][hjørnets]	125	41222(77215)
Fløj!	[fløj!]	110*43	[40705,s43]

Fig. 2: Excerpt from “Ved Frokosten” by Johs.V. Jensen (1905), “Gud ske Lov for Sofahjørnets Fløj!” (literal translation: *God be praised for the-sofa-corner’s velvet!*). Notice that “gud ske lov” is a lexicalized multiword expression, alternatively spelled “gudskelov”

Efficient Lambek provers are readily available. In addition, CLINK has a range of strategies for disambiguation based on morphological congruence, syntactical coherence, context cues and statistics (frequency tables, tensor models). See Henrichsen (2023) for discussion. CLINK’s selection algorithm is highly parameterized, allowing user-defined combinations of rule based analysis and machine learning. Of course, many parts of CLINK are not specific to the current project.

We present some early performance metrics (cf. Bick 2023); however, our intended contribution lies mainly with the generalized lexicological scope (the morph) and its associated representation (the template).⁵

We are currently developing COR.TALE, a COR₁ compatible database with phonetic information suitable for TTS. COR.TALE (due in 2025) will allow phonetic output along with the **cat** and **sem** annotations (with very little runtime/footprint penalty thanks to the CG-style compositionality).

5. Concluding Remarks

COR₁ and CLINK’s primary target group is the NLP developers. As many of these are neither trained linguists nor L1 speakers of Danish, it was imperative for the CLINK project to employ a lexicological approach with as little language specific grammar as possible. CG’s minimalistic rule base thus came in handy. As a corollary, the CG base also relieves the morphologist of the dilemma between a nomenclature derived entirely from the target language and one imported from a culturally dominant foreign language. Maybe the CLINK template, for these reasons, could find its use in foreign languages too. We are currently preparing a workshop for the smallest languages in

⁵ Measuring CLINK’s performance reliably is not trivial as we currently have no manually verified test corpora with CLINK annotation for **cat**, **sem** and **phon**. We are currently working on such a gold corpus, aiming at 10 mio. manually verified CLINK-annotated tokens extracted from DaGW (Derczynski, 2021). Preliminary tests suggest that CLINK’s performance is comparable to the Danish state-of-the-art (Asmussen, 2014; Bick, 2023), however the output of course being richer (including **sem** and **phon**).

the Nordic region (e.g., Greenlandic, Faroese, Samic languages) exploring CLINK as a basis for NLP; contact the author for more info.

In the Danish NLP sector (natural language processing), text preprocessors of all sorts (tokenizers, PoS-taggers, lemmatizers, classifiers) are being developed independently for TTS (text-to-speech), MT, word processors, chatbots and so on (Kirschmeier, 2019). We suggest a joint venture and invite all computational linguists to contribute to ordregister.dk with fresh COR-dictionaries (e.g., for specialized areas of expertise), parser modules, COR₁ related improvements and discussion. Above all, CLINK is intended as a framework for cross-organizational co-operations (such as Nimb, 2022; Bick, 2023) and friendly competition.

References

- Asmussen, J. (2014). *Design of the ePOS tagger*. Technical report, DK-CLARIN. korpus.dsl.dk/clarin/corpus-doc/pos-design.pdf
- Bick, E. (2023). Linking Danish Parser Output to a Central Word Repository. In: *Proceedings of KONVENS-19*. ACL (pp. 93–101).
- Derczynski, L. et al. (2021). The Danish gigaword corpus. In: S. Dobnik et al. (Eds.), *NEALT Proceedings Series No. 45* (pp. 413–421).
- Dideriksen, C. et al. (2022). Det Centrale Ordregister. *Nyt fra Sprognet 2022* (online only, <https://dsn.dk/nyt-fra-sprognet>). ISSN 2446-3124.
- Haspelmath, M. (2020). The morph as a minimal linguistic form. *Morphology*, 30, 117–134. <https://doi.org/10.1007/s11525-020-09355-5>
- Henrichsen, P. J. (2024). Tekstordet som grammatisk domæne. In J. Heegård (Ed.), *Ny forskning i grammatik* 31 (pp. 53–70).
- Jensen, J. V. (1906). *Digte 1906*. Gyldendalske Boghandel.
- Henrichsen, P. J. (2023). Det Centrale Ordregister. In: L. Holmer et al. (Eds.), *Nordiska studier i lexikografi* 16 (pp. 113–126).
- Jervelund, A. Å. et al. (2012). *Retskrivningsordbogen*. 4th edition. Danish Language Council Press (legal authority: cf. *Retsinformation*. ACT 332 14/05/1997, <https://retsinformation.dk>).
- Kirschmeier, S. et al. (2019). *Dansk sprogteknologi i verdensklasse*. Report commissioned by Ministry of Culture. ISBN 978-87-89410-77-7.
- Morill, G. V. (2010). *Categorical Grammar: Logical Syntax, Semantics, and Processing*. Oxford Linguistics.
- Nimb, S. et al. (2022). COR.SEM – den semantiske del af Det Centrale OrdRegister (COR). *Lexico Nordic*, 29, 73–95.

Wells, J. C. (1997). SAMPA computer readable phonetic alphabet. In: D. Gibbon et al. (Eds.), *Handbook of Standards and Resources for Spoken Language Systems*. Mouton de Gruyter. Part IV, section B.

Widmann, T. (2024). The Central Word Register of the Danish Language. In M. Medved, et al. (Eds.), *Proceedings of eLex 2023* (pp. 91–103).

Wood, Mary McGee (1993). *Categorial Grammars*. Routledge.

Contact information

Peter Juel Henriksen
Danish Language Council
pjh@dsn.dk

František Kovařík, Vojtěch Kovář, and Marek Blahuš

ON RAPID ANNOTATION OF CZECH HEADWORDS

Analysing the First Tasks of Czech Dictionary Express

Abstract Czech Dictionary Express has been introduced as a project of a semi-automatically made dictionary of the Czech language. (Kovařík, 2023) The Dictionary Express method (formerly known as *rapid dictionaries*) has been used for several different languages (Baisa et al., 2019; Blahuš et al., 2023). In this paper, we analyse the automatic and manual tools used in Czech Dictionary Express and inspect the statistical and qualitative data such tools provide. As the first task of the project – the headword annotation – comes to an end, we examine some opportunities and difficulties of the method used, as well as the data acquired in the process.

Keywords dictionary drafting; post-editing lexicography; corpus annotation; semi-automatic dictionary making; Dictionary Express; Czech

1. Introduction

Czech Dictionary Express (hereinafter referred to as CDE) is a project for creating a Czech dictionary from scratch. It uses a list of headwords generated automatically from large web corpora. This list is subsequently manually checked and revised by a team of Czech native speakers (the *editors*), who also inspect word forms and help disambiguate word senses in the later tasks. The process is controlled by supervisors (the *coordinators*), trained linguists and lexicographers.

One of the goals of CDE is to analyse the rapid dictionary-making method Dictionary Express (DE) used for several different languages. (Baisa et al., 2019; Blahuš et al., 2023) In the last 50 years, there has been a great effort in automating the dictionary-making process, so lexicographers can focus on more challenging and interesting tasks and linguistic phenomena. (Rundell & Kilgariff, 2011) Dictionary Express projects incorporate tools that help simplify dictionary making. Their aim is the possibility to create a dictionary of any natural language with a reasonably big corpus in a relatively short time span (ideally within a year) with a team of native speakers (not professional linguists, with secondary education), supervised by professional linguists and lexicographers who don't need to speak the described language (in the case of CDE, they do). (Further information on the DE methodology can be found on the <https://dictionary.express/> website. (DE, 2023))

This paper is to show the process of the headword annotation (the first manual task of the method) and interesting data gathered in the process of acquiring Czech headwords to make conclusions about the advantages and/or challenges of the semi-automatic approach.

2. Headword Annotation Task

The lexicon of the dictionary data was acquired from *Czech Web* (also known as *csTenTen*), a large Czech corpus consisting of texts from three large Czech web corpora: *Czech Web 2012*, *2017* and *2019*. (Suchomel, 2018) The texts were deduplicated and parts of the corpora were discarded for containing spam and automatically translated nonsensical texts. (Further information on *Czech Web* corpora family can be found on the web page. (cst, 2024)) The corpus was automatically lemmatized using the tools *Majka* and *desamb* (both recognized tools for the Czech language working with the Czech attributive tagset described by Jakubiček & Kovář, 2011). (Šmerk, 2008; 2009)

After the automatic acquisition, the project moved to the headword annotation task. In this section, we look closely at the annotation process and discuss interesting linguistic phenomena that have been discovered during the annotation process, mostly associated with a subjective view of word form, lemma form, and language standard.

2.1 Headword Annotation

The annotator team consists of 8 editors. 100,000 most frequent headword candidates (supposed lemmas with POS tags, e.g., *místnost-noun*) are annotated with a flag from a list of possible flags (see diagram in Figure 1): The word can be a non-word (the **I don't know** flag), a proper but **not Czech** word, a Czech but **non-standard** word, a standard Czech word but **not a lemma**, a standard Czech lemma but incorrectly POS-tagged (**wrong POS**), or it can be accepted as a **proper name** or a common word (the **OK** flag).

Before the task had started, each annotator was presented with an online course on lemmata and POS-tags, presented on the English language. They also attended a workshop where they learned about the task and some specific situations that might occur. With the help of an annotation manual, they then trained for the task on a small batch of word.

While working, the annotator sees a diagram (Figure 1) similar to the one used in the Ukrainian dictionary project. (Blahuš et al., 2023) The flag-assignment can be done using a mouse, but using a keyboard is faster and is preferred. The diagram shows the flags' colour coding and a button for the keyboard assignment of each flag.

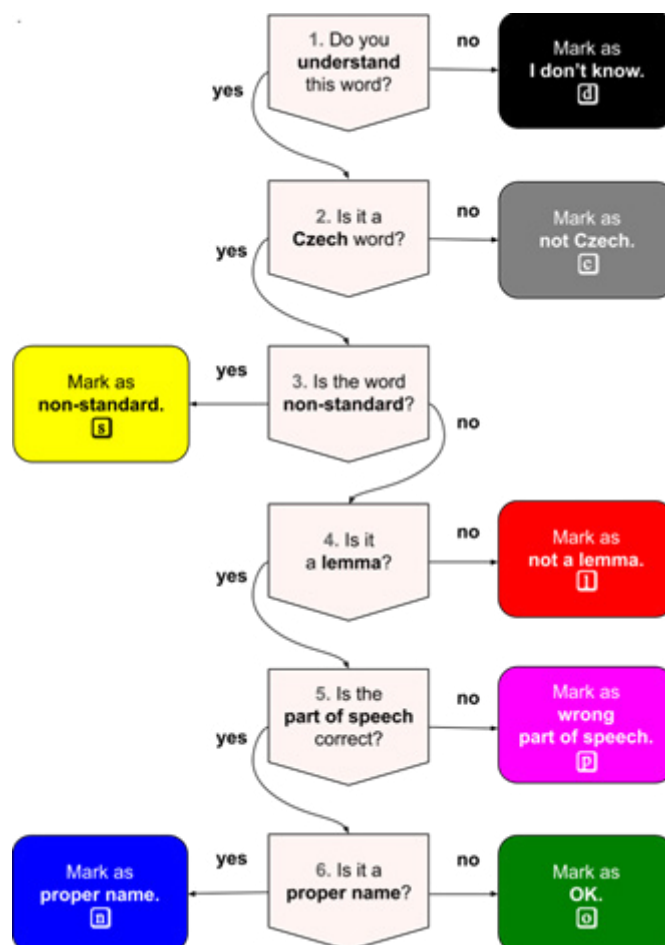


Fig. 1: Key to attributing flags to headword candidates.

The annotators didn't know the context of any particular headword and were advised against searching the words on the internet or in existing dictionaries. This made the process faster and prevented copying data from existing dictionaries.

For the annotation, the headword list is ordered alphabetically. The reason for this is that the annotator sees similar words close together, and can easily identify which words from the group are correct/incorrect.

2.2 Disagreement Patterns

Headwords are annotated at least twice. The 15,000 most frequent words have been annotated three times and 1,000 words have been annotated eight times, once by each annotator.

After analysing the data from the first 15,000 annotated headwords, it is possible to notice some rough patterns. Especially disagreement between two or more different annotators gives insight into what problems we should concentrate on.

Some level of disagreement is suppressed by applying **the presumption of correctness** (i.e., "if the headword suggested by the lemmatizer/tagger can be

considered correct, it should”) and pointing out expected problems (as described in Kovařík, 2023). The presumption of correctness is an important principle. The CDE is centred around swift dictionary making and language phenomena that make sense for the average educated native speaker, not around “linguistic purity”. The task of the annotator is to correct the assumption of the lemmatizer/tagger only if it is transparently wrong, in contradiction with the target language.

However, not all difficulties can be solved like this and after some time we could see new recurring problems:

- Some annotators marked the **diminutives** with the *not a lemma* flag. Diminutives usually go as separate lemmas in Czech and often have a different meaning than just a simple diminution of the non-diminished word. Diminutives should thus be treated as separate lemmas and not forms of their non-diminished relatives.
- The manual emphasises that **short forms of adjectives** (e.g., *zdráv*) should be marked with the *not a lemma* flag. Only the long (the so-called “složený”, composed) form of adjectives (e.g., *zdravý*) should be accepted as a lemma. Despite this, annotators continued to mark the short forms of adjectives *OK*.
- As expected, controversy arises over the **view of the standard** (a classification of a word as “standard” or “non-standard”). This includes disagreement over the **dialectisms** (e.g., *tož*), **past forms** (mostly the difference between the suffix *-ismus* and more recently accepted *-izmus*; some words ending with *-ismus* are viewed more historical than others by different people) and more (e.g., the difference in *o/ó*, like in the word *salon/salón*). Most of these words are marked *non-standard* by one annotator and *OK* by another.
- For some words, it’s difficult to **decide the POS**. For example, the word *zima* is a noun but can also in some sentences be considered an adverb: *Je mi zima*. Another problem is the nominalised adjectives (nouns derived from adjectives) that, with no context, can be mistaken for pure adjectives. Examples of these are *zlatá-noun*, *zelená-noun*, *výborná-noun*, *známý-noun*, *ženská-noun*, *zraněný-noun*. All of these should be accepted (because of the presumption of correctness). Controversy also arises over words that are considered pronouns or numerals but are used as other parts of speech in a sentence (e.g., *víckrát* could be a numeral or an adverb).
- Does the **non-negated form** exist? Is it used in Czech? This question arises frequently, e.g., for the words *překonatelný*, *přetržitý*, *vyhnutelný*, *vyzpytatelný*, *zbytný* or *zvykle*.
- In many cases, one annotator marked a headword *OK* and another *not Czech*. This disagreement arises most often in the cases of **words coming from English** that are only partially accepted by the Czech native speakers (*link*, *market*, *teenager* etc.), or in the cases of the only partially accepted

proper names (*Juraj, Trump, Times* etc.) or name forms (*Jozef* which is more a Slovak form than Czech etc.).

- Deciding whether a word coming from other languages is or isn't part of target language is a difficult linguistic task which many times doesn't have a clear answer. However, the Dictionary Express method can at least provide useful data on how the native speakers approach a given (foreign) word.

2.3 Inter-Annotator Agreement and Annotation Statistics

One of the aims of the CDE project is to study the view of the language from the perspectives of annotators. This means not only defining the language-specific problems but also observing the annotation statistics.

Statistics provided by the frequency-acceptance relationship show that the less frequent a word is, the less likely it is going to be a proper Czech word (Figure 2). Hence, there is more or less a direct proportion between word frequency and chance of accepting the word.

Figure 2 also shows how the results of each annotator differ. Less frequent words show a greater difference between some annotation statistics. The outstanding results (the two lowest) were checked manually by the coordinators of the project, and are caused mostly by the alphabetical order of annotated headwords. They are connected to the quality of the particular annotations.

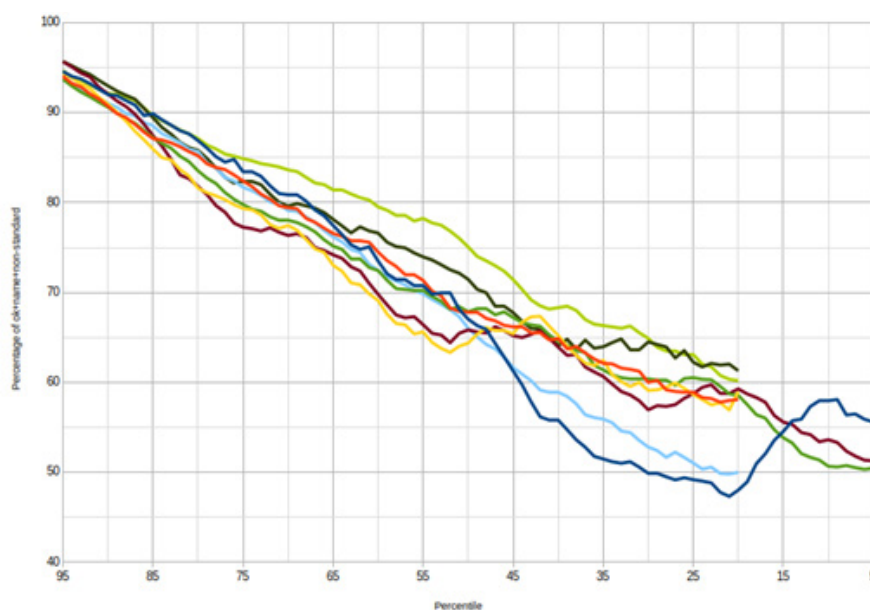


Fig. 2: The less frequent the headword is in the corpus, the less likely it is going to be a proper Czech headword. Vertical axis presents how many words have been marked as *OK*, *proper name* or *non-standard*. Horizontal axis represents frequency (most frequent words on the left) divided by percentile. Each coloured line represents an annotator. Calculated from 95,000 annotated headwords.

Especially useful is the number of cases when the annotators' opinions are the same and cases when they are different – the so-called inter-annotator agreement. Tables 1 and 2 show statistics of 205,926 headword annotations.

- **Table 1** demonstrates, how many times annotators agreed upon a flag marked to a headword.
 - In the left column: The first number is the maximum times the annotators agreed upon a flag for a headword. The second number is the times the headwords have been annotated. Hence, “4/8” means “headwords for which a maximum of 4 annotators agreed on a flag and which have been annotated 8 times”.
 - In the right column are the counts of these headwords.
- **Table 2** shows the statistics of flags marked to words with an inter-annotator agreement greater than 50 %. E.g., 47,593 headwords were mostly marked OK.

Maximum agreements / times annotated	Number of headwords
1/1:	15,000
1/2:	12,784
2/2:	42,290
1/3:	782
2/3:	3,813
3/3:	18,331
2/4:	18
3/4:	78
4/4:	904
3/8:	14
4/8:	57
5/8:	102
6/8:	112
7/8:	194
8/8:	521

Table 1. Inter-annotator agreements

OK	47,593
name	5,996
not a lemma	4,510
non-standard	527
wrong POS	690
I don't understand	3,474
not Czech	1870

Table 2. Flag statistics (more than 50 % agreement)

3. Annotation Revisions

In the annotation task of the project, more than 100,000 headwords have been annotated with two or more flags. Some of the headwords still need to be revised: those marked with two or more different flags and those marked *non-standard*, *not a lemma* and *wrong POS*. A group of experienced editors is selected to examine the annotations and decide, as objectively as possible, which flag is appropriate and how to approach partially unaccepted headwords and non-standard forms. We call this process *the revision(s)* and the editors charged with the task *the inspectors*.

The inspector goes through a list of words annotated in the headword annotation task. This list consists of words mostly annotated with the *non-standard*, *not a lemma* or *wrong POS* flags and of words with a combination of different flags. Words annotated

only with the *I don't know* and/or *not Czech* flags, only with the *proper name* flag or only with the *OK* flag are not included. This means only a fraction of the words (approximately 25,000 of the 100,000 in CDE) need to be revised.

The screenshot shows the Lexonomy interface for revising a Czech headword. At the top, the word 'Bříza' is labeled as a NOUN. Below this, there are two red buttons labeled 'not a lemma?' and 'opinions: not a lemma not a lemma proper name'. The main section is titled 'CORRECT HEADWORD SHOULD BE:' and contains two rows of input fields. The first row has 'lemma: bříza' and 'PoS: noun', with a 'proper name?' checkbox that is unchecked. The second row has 'lemma: Bříza' and 'PoS: noun', with a 'proper name?' checkbox that is checked. Each row has a 'DELETE' button with a trash icon. Below these rows is an 'ADD MORE? +' button. At the bottom, there are three radio buttons: 'I DO NOT UNDERSTAND THIS WORD', 'THIS IS NOT A CZECH WORD', and 'THIS HEADWORD IS CORRECT'. Below the radio buttons is a section titled 'EXAMPLES' with three numbered sentences: 1. 'Bříza má obvykle rovná vlákna s jemnou strukturou.' 2. 'Situace bude rozhodovat pan Bříza ze své kanceláře.' 3. 'Domácí okamžitě začali útočit ale v cestě stál Bříza.'

Fig. 3: Example of a revision entry.

Figure 3 shows the Lexonomy interface (Měchura, 2017), where the revisions get done. The inspector looks at a headword and its previous annotations. The annotations direct them to a more objective approach to the given headword than in the previous task. At the inspector's disposal is also the context typical for the headword (demonstrated as a list of Good Dictionary Examples (Rychlý et al., 2008).

Based on the provided data and knowledge of Czech as their native language, the inspector decides that:

1. the headword is slightly incorrect or has two or more different correct forms, and enters the correct form of the headword with the option of marking it a proper name;
2. OR the headword's lemma is not a proper Czech lemma;
3. OR they don't know the headword, in which case the headword is marked as a non-word;
4. OR the headword is correct.

If the headword is a non-standard form of a formally similar standard Czech headword (e.g., *čtyry-numeral*, it should be corrected to the standard form (e.g., *čtyři-numeral*).

If the headword is an incorrect or non-standard union of two separate words (e.g.,

zachvilka-noun in Czech), it should be corrected to the correctly used phrase of two or more words (e.g., *za chvilku – adverb* in Czech).

The revision is a significantly more complicated task than the headword annotation because it expects the inspector to view the headwords more objectively. Thus, the inspectors were chosen from the editors who had more experience with annotation and who proved more consistent while annotating headwords.

4. Subsequent Tasks

The creation of an annotated and revised headword list is only the first task of the methodology of creating a full dictionary of a language. Subsequent tasks of CDE are being prepared and executed. These include:

- annotation of inflected forms of nouns, adjectives, verbs, adverbs and other, language-specific categories;
- word sense disambiguation, translation to English, choosing the right dictionary examples etc.;
- recording of word pronunciation (the only manual-only task of DE).

5. Conclusion

In this paper, we described the annotation and revision tasks of the Dictionary Express method for dictionary-making and demonstrated it on the ongoing Czech Dictionary Express project. A headword list is created using large, automatically lemmatized corpora. In the annotation task, the annotators go through this list and assign flags such as *not Czech*, *non-standard*, *not a lemma* or *OK*. In the revision task, the inspectors go through the list again, look at the assigned flags and decide more objectively which headwords should remain in the list and in what form.

By the time of writing, almost 100,000 headwords have been assigned at least two flags. This provides interesting data for an insight into the language and its perception. Some of the data have been discussed in the paper.

A list of headwords has been created for the dictionary. The Czech Dictionary Express project continues with tasks concerning inflected forms. Later on, word senses will be assigned to headwords through a combination of automated tools and manual work, as well as translation to another language (English), dictionary examples and other parts of the dictionary.

References

DE, 2023. (2023). *Dictionary Express – automated dictionary generation*. Retrieved April 17, 2024, from <https://dictionary.express/>. Accessed:

cst, 2024. (2024). *csTenTen – Czech corpus from the web*. Retrieved April 17, 2024, from <https://www.sketchengine.eu/cstenten-czech-corpus/>

Baisa, V., Blahus, M., Cukr, M., Herman, O., Jakubíček, M., Kovář, V., Medved', M., Mechura, M., Rychlý, P., & Suchomel, V. (2019). Automating dictionary production: a tagalog-english-korean dictionary from scratch. In *Proceedings of the 6th Biennial Conference on Electronic Lexicography* (pp. 805–818). Brno, Czech Republic. Lexical Computing CZ s.r.o.

Blahuš, M., Cukr, M., Herman, O., Jakubíček, M., Kovář, V., Kraus, J., Medved', M., & Ohlidalová, V. (2023). Rapid Ukrainian-English Dictionary Creation Using Post-Edited Corpus Data. In M. Medved', & M. Mechura (Eds.), *Electronic lexicography in the 21st century (eLex 2023): Invisible Lexicography. Proceedings of the eLex 2023 conference*, Brno, Czech Republic. Lexical Computing CZ s.r.o.

Jakubíček, M., Kovář, V., & Šmerk, P. (2011). Czech morphological tagset revisited. In R. Horák (Ed.), *Proceedings of Recent Advances in Slavonic Natural Language Processing 2011* (pp. 29–42). Brno. Tribun EU.

Kovařík, F. (2023). Semi-automatic dictionary creation for czech. *Recent Advances in Slavonic Natural Language Processing (RASLAN 2023)*, 17.

Měchura, M. (2017). Introducing Lexonomy: An Open-source Dictionary Writings and Publishing System. In I. Kosem, C. Tiberius, et al. (Eds.), *Electronic lexicography in the 21st century: Proceedings of eLex 2017 conference*. Lexical Computing.

Rundell, M., & Kilgarrieff, A. (2011). Automating the creation of dictionaries: Where will it all end? In F. Meunier, S. De Cock, G. Gilquin, & M. Paquot (Eds.), *A Taste for Corpora: In honour of Sylviane Granger. Studies in Corpus Linguistics*, 45 (pp. 257–282).

Rychlý, P., Husák, M., Kilgarrieff, A., Rundell, M., & McAdam, K. (2008). GDEX: Automatically finding good dictionary examples in a corpus. In E. Bernal, & J. DeCesaris (Eds.), *Proceedings of the XIII EURALEX International Congress* (pp. 425–432). Barcelona. Institut Universitari de Lingüística Aplicada.

Suchomel, V. (2018). csTenTen17, a Recent Czech Web Corpus. In A. Horák, & A. Rambousek (Eds.), *Proceedings of the Twelfth Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2018* (pp. 111–123). Brno. Tribun EU.

Šmerk, P. (2008). *Towards Morphological Disambiguation of Czech*. Ph. D. thesis proposals, Faculty of Informatics, Masaryk University.

Šmerk, P. (2009). Fast Morphological Analysis of Czech. In *Proceedings of the Raslan Workshop 2009*, Brno. Masarykova univerzita

Irene Renau, Rogelio Nazar, and Daniel Mora

TOWARDS THE AUTOMATIC GENERATION OF A PATTERN-BASED DICTIONARY OF SPANISH VERBS

Abstract Corpus Pattern Analysis, CPA (Hanks, 2004a; 2004b; 2013), is a technique for identifying local semantic and syntactic information of a word and mapping it to its meanings. In verbs, it consists basically of the argument structure labelled with semantic types for each argument. CPA is used in several dictionary projects and allows systematic corpus analysis; however, it is extremely time-consuming. In this paper, we present a method for the automatic pattern identification of Spanish verbs in corpora. We used a syntactic parser for dependency analysis (Stanza), applied a named entity recognition (NER) tagger from the Flair NLP framework for NER and, for common nouns, we implemented a semantic tagger and a word sense disambiguation method, both created for the task. All resources were combined to extract CPA verb patterns. The method performs better than previous attempts and can contribute to a more efficient pattern-based lexicography.

Keywords argument structure; Corpus Pattern Analysis; pattern-based lexicography; semantic tagging; word sense disambiguation

1. Introduction

In this paper, we present a method for verb pattern recognition in corpora. For this particular experiment, we work with Spanish transitive verbs and use a series of strategies that, combined, can extract syntacto-semantic structures from discourse. A pattern can be defined as a piece of phraseology which is recurrent in discourse and can be mapped to a particular meaning of a verb (Hanks, 2004a; 2004b; 2013; Sinclair, 2004). For example, if someone says *We love Patrick*, it means having ‘strong positive feelings for Patrick’, while saying *We love parties* means ‘liking parties very much’. The idea is that the pattern *[[Human]] love [[Human]]* is the pattern beneath the first sentence, and *[[Human]] love [[Event]]* is the one beneath the second one. Corpus Pattern Analysis, CPA (Hanks, 2004a; 2004b; 2013), offers a systematic way of analysing these patterns in corpus data and has been used in many dictionary projects of different languages. While this technique is highly sound both theoretically and methodologically, it is also very time-consuming. Our goal then is trying to find a way for the teams to extract these patterns from a corpus more efficiently.

In this article, we first present a state of the art of CPA (Section 2): its theoretical and methodological background, a summary of dictionaries and lexical databases using CPA to analyse corpus data, and a mention of some previous attempts to automate pattern recognition. We then explain the method (Section 3), dividing it into syntactic and semantic analysis, and offer an evaluation of results (Section 4). We finish with some conclusions and limitations of the experiment, which point to future work (Section 5).

2. Corpus Pattern Analysis and Pattern-Based Lexicography: State of the Art

2.1 Theoretical and Methodological Background

CPA is a method to disambiguate words in context (Hanks, 2004a; 2004b: 2013; Hanks & Ježek, 2008; Hanks & Pustejovsky, 2005; cf. also Grefenstette & Hanks, 2023). It is described by Hanks (2004a, p. 87) as a “technique for mapping meaning onto words in text”. It is based on a vast body of work considering that words do not have meaning in isolation, but most likely a meaning potential that is activated in a certain context of usage. Sinclair (1991; 2004), the closest and important antecedent of CPA, proposes that the problem of meaning must be tackled from a syntagmatic perspective, being corpus analysis of paramount importance “to observe recurrent patterns of language” (Sinclair, 2004, p. 140).

CPA develops the Sinclairian notion of pattern into a fully semi-automatic method of corpus analysis. Context of analysis is restricted to local context, which “is usually sufficient to assign a specific sense to a word and to distinguish one sense from another” (Hanks & Pustejovsky, 2005, p. 64). A pattern, thus, is broadly defined in CPA as “a semantically motivated and recurrent piece of phraseology” (Ježek & Hanks, 2010, p. 8). As we will show in sections 2.2 and 2.3, CPA is used mainly for verbs. In verbs, a brief description of a pattern is given by the acronym ‘T-PAS’ (Ježek et al., 2014; see Section 2.2): *Typed Predicate Argument Structure*. Indeed, a pattern consists of the argument structure of the verb and the semantic typing of each argument. Some additional syntactic information is added when required, such as pronouns for pronominal verbs (e.g., in Romance languages) or prepositions in some type of complements.

Semantic types used to label each argument are organised in an ontology suitable for everyday language, containing basic conceptual categories (Hanks & Pustejovsky, 2005; Ježek & Hanks, 2010). Semantic types in this ontology must be specific enough to distinguish one pattern from the other, especially when the syntactic structure is identical. However, categories are not fine-grained to the point that they are useless for the task, e.g., for CPA, a category such as *cooking apple* would be unnecessary.

Figure 1 shows an example of CPA for a verb extracted from the *Pattern Dictionary of English Verbs*, PDEV (Hanks & Pustejovsky, 2005).

love	
1	[[Human 1 Deity Animal]] love [[Human 2 Deity Animal]] [[Human 1]] has strong affectionate feelings for or cares very much about [[Human 2 Deity Animal]]
2	[[Human Animal]] love [[Inanimate Abstract_Entity]] [[Human Animal]] very much likes [[Inanimate Abstract_Entity]]
3	[[Human]] love [[Event]] [to+INF] [[Human]] greatly enjoys [[Event]] {to/INF [V]}
4	[[Human]] would love [to+INF] [[Human]] would very much like {to/INF [V]}
5	[[Human Animal]] love (it) [THAT] [WH+] [[Human Animal]] is very happy ([[that-CLAUSE WH-CLAUSE]])

Fig. 1: Entry for *to love* in the PDEV [last access: 20/5/2024]

2.2 CPA Projects

There are currently multiple projects using CPA for lexical analysis and lexicography in different languages. The pioneer project is the PDEV, developed under the direction of Patrick Hanks, mostly at the University of Wolverhampton. In its public version, it includes at the moment 1,691 verb entries, each one containing a list of the normal patterns of verbs and the description of the meaning–implicature–mapped onto each pattern. The dictionary is organised as a database in which users can also look for the semantic types of the patterns. For example, the category *[[Artefact]]* can be searched and, as a result, we obtain the list of all verbs in which this category is used at least in one of the patterns. The CPA Ontology containing the organisation of all semantic types is also provided in one of the sections of the webpage.

PDEV has been the foundational project and the inspiration for an array of other initiatives. Especially similar to PDEV are T-PAS, CROATPAS and Verbario, all lexical databases suitable to be used for downstream NLP tasks. T-PAS (Ježek et al., 2014) is a database of Italian verbs. It contains 1,000 verbs with different degrees of polysemy and can be downloaded with a Creative Commons license. CROATPAS (Marini & Ježek, 2019; Marini, 2022) is developed following T-PAS and is used to analyse the polysemy of Croatian verbs. At the moment, it contains 180 verbs and a total of 683 patterns distributed in the entries. CROATPAS shows the feasibility of using CPA for under-resourced languages. The Verbario database (Renau et al., 2019) contains around 200 verbs. Each section of the verb entry contains the pattern, the meaning (or implicature of the pattern), and a link to all concordances that were analysed to extract the patterns from corpus (around 250-1,000 concordances per verb are typically analysed in Verbario, depending on the complexity of the verb). Concordances are linked to each pattern and, in each entry, patterns are displayed with numbers. In cases of inchoative verbs, the alternance is shown by adding a letter to the number.

CPA is also used in different learners' dictionary projects. They have in common that they use CPA for corpus analysis, but they adapt the formulation of the pattern to learner's needs and to the specific project requirements, which implies different types of modifications and simplifications. It is the case of an advanced learners' dictionary of Italian currently in progress (DiMuccio-Failla & Giacomini, 2017; 2022), The *Diccionario de aprendizaje del español como lengua extranjera*, DAELE, a proposal for a learners' dictionary of Spanish (Battaner & Renau, 2011), or the DSELE, a proposal for pronominal verbs in Spanish in a learners' dictionary (Renau, 2012). Finally, the *Woordcombinaties* (Colman & Tiberius, 2018) is an ongoing project for a Dutch learners' dictionary. In all these cases, CPA is used for corpus analysis and, after creating the patterns, they are adapted to the dictionaries.

2.3 Computer Methods for CPA

All dictionaries and databases described in sections 2.2 and 2.3 have in common that they use semi-automatic procedures, mainly different Sketch Engine tools (Kilgarrieff et al., 2014). However, a large amount of work is still done manually and it is extremely time-consuming. There have been some proposals in this way since the beginnings of the PDEV—some of them are summarised by Giacomini and DiMuccio-Failla (2019). Pustejovsky, Hanks & Rumshisky (2004) made a preliminary attempt for automatic pattern acquisition incorporating tokenisation, POS tagging, lemmatisation and dependency analysis to corpus sentences, observing if results matched the manual analysis. In this study, semantic tagging seems to be a human operation nonetheless. Baisa et al. (2015) describe a SemEval shared task based on CPA, in which three sub-tasks were proposed, working with English data only: 1) CPA parsing (syntactic and semantic) of corpus sentences, 2) CPA clustering (of corpus sentences according to their similarities) and 3) CPA lexicography, which involved pattern building according to previous parsing and clustering. Teams report results on sub-tasks 1 and 2 but not for 3, and no team participated in more than one task.

In the case of the Verbario project, different efforts have been made to automate the procedure of pattern induction for Spanish verbs covering all these tasks, from parsing to pattern building (Renau et al., 2019). The following steps are necessary for replicating the manual procedure with computational methods: 1) Extracting corpus concordances, 2) Using a parser for dependency analysis in order to identify the main verb and its argument structure, 3) Semantic tagging of the arguments, that is, to identify the category of the head of each constituent of the sentence playing the role of argument, and 4) Pattern building. The attempts made in the Verbario project following these steps were promising but still left ample room for improvement (about 60% of the patterns were acceptable). The main problem were the difficulties of the semantic tagging and the lack of a method to deal with polysemy in the hypernymy links of the target nouns (Nazar & Renau, 2016).

3. Methods

In this section, we propose an improvement of the method used in the Verbario project (shown in Section 2.3) to automate tasks of corpus analysis and pattern building. The basic approach remains the same, as we imitate the manual process following the four steps already mentioned. However, we made adjustments in all steps to improve precision, especially concerning semantic tagging. For this experiment, we dealt only with transitive verbs and, specifically, only those with two arguments: subject and direct object. A summary of the method is shown in Figure 3.

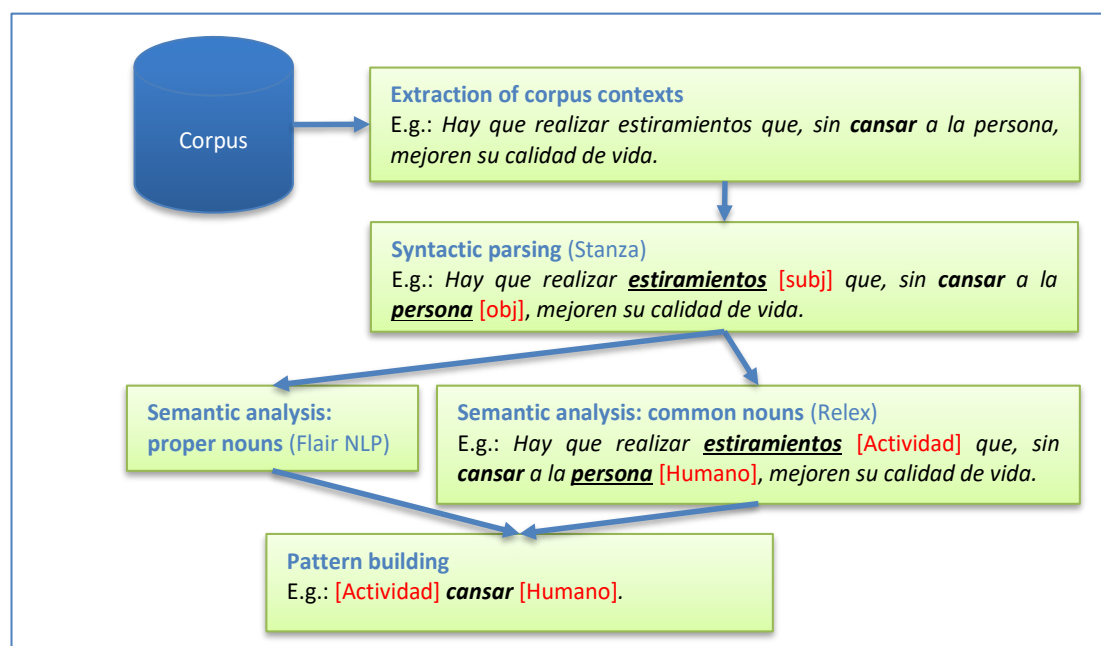


Fig. 3: Synthesis of the method

3.1 Corpus and Data Processing

We used the Spanish section of the Opus Corpus (Tiedemann, 2009), consisting of 25 billion words of texts from a variety of genres and sources. For the analysis we excluded very long sentences which may contain complex clauses. Our idea was to try to simplify the task for the syntactic parser, as previous attempts of parsing without filtering led to unsatisfactory precision (see Section 4.1). Also, we used the esTenTen corpus (Kilgarriff & Renau, 2013) to extract word association coefficients (see Section 3.3.3).

3.2 Syntactic Parsing

The syntactic parsing was carried out using the Stanza library (Qi et al., 2020), as it is slightly better compared to UDPipe 2.0 (Straka, 2018) in the Spanish-AnCora corpus. Stanza is made up of a neural network pipeline that contains several processors that take care of different text processing tasks, and annotates using universal dependencies and universal features. In this case, we used its lemmatisation module, part-of-speech / morphological tagger, and dependency parsing Spanish processors.

In order to match up both the syntactic and the semantic parsing at token level, the input was first tokenised by the semantic parsing model, and then fed to Stanza. The main focus was on extracting the basic argument structure (subject and direct object), considering transitive verbs only. For simplicity, if a syntactic argument of a verb is a noun phrase, we only consider the head noun, and if two sentences are joined by a conjunction, we retain the structure of the first sentence. The output, therefore, is each token annotated at the syntactic level, and the main tree structure (subject, root, direct object) of the sentence. For example, a sentence such as *Nepal ha presentado*

periódicamente informes al Comité ‘Nepal has periodically submitted reports to the Committee’ has *presentar(Nepal, informe)* as output.

3.3 Semantic Parsing

3.3.1 Proper Nouns

We identified proper nouns using a NER tagger from the Flair NLP framework (Akbić et al., 2019a), specifically, the ‘ner-multi-fast’ model (Akbić et al., 2019b), available through the HuggingFace library (Wolf et al., 2020). This model was preferred over other Spanish NER taggers given its trade-off between processing speed and accuracy. The ‘ner-multi-fast’ model is trained on six different languages at a character level, including Spanish. In our experiment, the Flair model receives tokenised text from Stanza and, for each token, it outputs its NER tag (i.e., Person, Location, Organization, Miscellaneous) in BIO format, e.g., *El texto fue escrito por el poeta de Weimar[LOC] Salomon[B-PER] Franck[E-PER]* ‘The text was written by Weimar’s poet Salomon Franck’ (LOC: location; PER: person).

3.3.2 Common Nouns

We populated a taxonomy of common nouns using a general Spanish dictionary and a word sense disambiguation (WSD) system. The dictionary is used to obtain the hypernym from the definition of each noun, and the WSD system is used to concatenate large hyponym-hypernym sequences in a coherent manner. Without a WSD system, the hypernym chains would quickly derail by the effect of polysemy, as already noted by Amsler (1981). For instance, from the dictionary it is possible to determine that a *ñacaratiá* is a kind of *árbol* ‘tree’, and then that an *árbol* is a kind of *planta* ‘plant’. But then *planta* is a polysemous word, like in English, where we have for instance ‘industrial installation’ that is not a ‘living organism’. Thus, the system must figure out what sense of the word *planta* is said to be the hypernym of *árbol*. Furthermore, a WSD is also needed later during the task of semantic tagging of the analysed text, as there will be frequent encounters with polysemous nouns in the texts to be tagged, words that have to be disambiguated before assigning them a semantic type.

The application of a WSD system for the development of a taxonomy means that one has to start from each noun in the lemma list of the dictionary and work the way up to some of the semantic types of the top-ontology. The only advantage of the application of the WSD system to taxonomy extraction over the typical case scenario is that one can use the inheritance mechanism, i.e., to use the text of the definitions retrieved from the lower links of the hypernym chain. The words in the definition of *ñacaratiá* and *árbol* are helpful to select the corresponding sense of *planta* because of the semantic similarity of the words in the definitions. In other words, we have a capital of keywords to distinguish between a ‘living organism’ and, say, an ‘industrial installation’.

As the dictionary, we used the Spanish Wiktionary, a useful resource seemingly underused in NLP. Inspired by an old tradition in computational linguistics (Chodorow et al., 1985; Guthrie et al., 1990, and many others), we designed a rule-based algorithm to extract the hypernyms from the definitions which is able to circumvent the different obstacles presented by the lack of uniformity and systematicity of this lexicographic resource. In most cases, hypernyms correspond to the first noun in the definition, as in *ñacaratiá*, defined as “Árbol frutal sudamericano...” ‘South American fruit *tree*...’. However, there are many exceptions and one has to avoid metalinguistic expressions (*type of...*, *kind of...*, *name of...*, etc.) and scientific names of animals and plants, among other information.

This rule-based method of taxonomy induction from a single dictionary is a significant departure from earlier attempts in the context of the Verbario project, which were based on statistical corpus analysis and did not adequately handle the cases of polysemy. We leave for future work the possibility of combining earlier ideas (e.g., Nazar & Renau, 2016) with the method now proposed.

3.3.3 Word Sense Disambiguation Methods

Regarding the WSD system, in Spanish the subject has been thoroughly investigated (Agirre et al., 2014; Bevilacqua et al., 2021, among others). However, to the best of our knowledge, there are scarce open-source implementations.

For the development of the WSD system, we again went to the basics, following the policy of first putting in place something simple but functional, that could be progressively developed. Starting from the first ideas by Lesk (1986), we conceptualise the task of WSD as follows: there is an input sentence that contains a polysemous noun to be disambiguated, we have a list of senses of such word in a dictionary, and we have to select the most likely sense on the basis of the associations we find between the context of the noun and the text of the definitions. Lesk’s (1986) first idea was to use the vocabulary intersection between both sets, but this is not very promising, as the chances of finding such an intersection are minimal.

3.3.3.1. Statistical WSD With Relex

Our first attempt was to redefine Lesk’s (1986) proposal and develop a statistical WSD method (*Relex*), which instead of the intersection tries to calculate a semantic similarity between definition and target noun. Of course, if any intersection, however unlikely, happens to occur, it is also considered. Most of the computation, thus, involves words that are not the same, but that are semantically related. We do not need a precise definition of semantic relatedness, and it could be very diverse. We say, for instance, that the nouns *cat* and *dog* are semantically related. Conceptually, they are co-hyponyms, but operationally, we considered that any pair of words is semantically related if they show a strong syntagmatic association in a large Spanish reference corpus, in our case, the EsTenTen. To compute these associations, we analysed a sample of the corpus (1/1,000) and produced tables of syntagmatic associations

between pairs of all kinds of words, i.e., independently of their grammatical category. The values of this table are word₁, word₂, total frequency of word₁, total frequency of word₂, the frequency of co-occurrence in the same sentences and an association measure. Relex then uses this table to compare the vocabulary of the target text and each definition at a time, collecting the values of the association of each pair of words. The values are summed for each sense.

For illustration, consider the case of the frequently used polysemous Spanish word *virus*. The dictionary lists two senses: one with *organismo* ‘organism’ as hypernym and another with *software* as hypernym. Given a certain context of usage, there will be a strong association between collocates such as *pacientes* ‘patients’ or *transmisión* ‘transmission’, etc. and the first meaning, while words such as *antivirus* ‘anti-virus’, *herramienta* ‘tool’ or *computadora* ‘computer’, etc. will be strongly associated to the second meaning.

3.3.3.2. LLM-Based WSD

We also developed another method with three alternative approaches that use MarIA (Gutiérrez-Fandiño et al., 2022), the most extensive RoBERTa-based large language model (LLM) trained on Spanish corpora. These three approaches followed common and novel strategies for WSD using LLM (Huang et al., 2019; Bevilacqua et al., 2021), making special use of the fill-mask task and cosine similarity.

The first approach (‘fill-cooc’) extends Lesk’s (1986) intuition by measuring the vocabulary intersection between a candidate’s definition and the definitions of 10 words generated by the model when doing fill-mask on the target word. Fill-mask is the task in which the model predicts the masked token within a sentence. For example, the target *ratón* ‘mouse’ from the sentence *Se pulsa el botón del ratón* ‘The mouse button is pressed’ is masked and fed to the model (i.e., “The [MASK] button is pressed”). It is expected that by expanding the sets of words, intersections between definitions are more likely to occur.

The second and third approaches use cosine similarity, a metric for measuring semantic relatedness between two vectors. The value ranges from -1 (very dissimilar) to 1 (very similar), with 0 indicating no relation. Word embeddings, generated by the LLM, are used to obtain these vectors. The second approach (‘fill-cosine’) vectorises each definition and the concatenation of the 10 generated words after applying fill-mask. Then, cosine similarity is measured between the candidate definition and the generated words, thus circumventing the direct intersection of words.

The third approach (‘target-cosine’) computes cosine similarity between the vectorised input sentence and each definition of the target word without doing fill-mask. This approach exploits contextualised word embeddings generated by the LLM. Contextualised word embeddings adjust based on the specific context in which the word appears. When applying cosine similarity (approach 2 and 3), the definition with the highest score is the predicted one.

3.4 Pattern Building

Once we have all the Opus Corpus fully tagged with the different layers we have been describing in sections 3.2 and 3.3 (i.e., POS-tags, syntactic parsing and semantic tags), we proceeded with the construction of CPA patterns. For each of a list of Spanish verbs, we extracted a maximum of 5,000 sentences that have them as the main verb of a simple transitive structure (we reserve for future work the possibility of using other structures). For each sentence, we collected the semantic type of both syntactic arguments of the verb and computed a frequency table of the types.

For illustration, consider the case of the verb *aburrir* ‘to bore’ used in contexts such as *Él nos aburrió* ‘He bored us’, *Esa gente me aburre* ‘This people bore me’, etc. Each of these sentences is recognised as an instance of the pattern *[[Humano]] aburrir [[Humano]]* ‘[[Human]] bore [[Human]]’. On the other hand, cases such as *El debate me aburre* ‘The debate bores me’, *Los trabajos le aburrían* ‘Jobs bored him/her’, etc. are instances of the pattern *[[Acción]] aburrir [[Humano]]* ‘[[Action]] bore [[Human]]’. In this way, as the analysis progresses and the number of sentences increases, so does the importance of each pattern in the frequency table. Once all the sentences of a given verb are analysed, we retain the *n* most frequent patterns.

The number of different patterns per verb will increase or decrease depending on the frequency of the verb but also depending on how general the semantic type is. We have retained this value as an execution parameter: if the system considers as semantic types only very abstract nouns (i.e., those placed in the higher levels of the taxonomy), then the number of different patterns decrease, as different sentences are lumped into the same pattern. On the contrary, using more specific (lower) semantic types splits the patterns in many. For instance, if instead of, say, *[[Vehicle]]*, we consider more abstract semantic types such as *[[Artefact]]*, then different artefacts other than vehicles will be placed under the same pattern. A systematic method to automatically optimise this level of abstraction is left for future work.

4. Results and Evaluation

4.1 Evaluation of the Syntactic Parser

For the evaluation of the syntactic parser (see Section 3.2), we drew two random samples from the Opus Corpus of 100 sentences each: the first one had no restrictions, i.e., all types of sentences had the same probability of being in the sample, and the second was limited to simple transitive sentences, i.e., no more than two arguments, no attachments and no subordination. Table 1 shows results of the evaluation of both samples.

Table 1: Figures of precision in percentage of syntactic parsing with Stanza

	Correct	Incorrect	Unanalysable
Unrestricted sample	50	32	18
Simplified sample	81	19	0

We limited our error analysis to those cases in which the system was unable to correctly identify the main verb and the arguments of the predicate structure. Unsurprisingly, these were more widespread in the case of the unrestricted sample, which contains larger and more syntactically complicated sentences. In the case of the simplified sample, there are still many errors. However, an 81% success rate is quite acceptable for our purposes, and it is only moderately lower than the results reported in the evaluation of Stanza with the Spanish-AnCora corpus. The 31% difference between both samples confirms that the decision to restrict the corpus was correct.

4.2 Evaluation of the Proper Noun Semantic Tagger

For the evaluation of the NER system (see Section 3.3.1), we proceeded with the same random sample of 100 sentences used in the previous section, specifically the unrestricted sample, and analysed in this case the identification and categorisation of the proper nouns. We manually identified a total of 112 proper nouns in the sample. All but two were correctly detected by the NER tagger (98% success). With regard to the semantic categorisation of the nouns, 94 of the 110 (85%) received the correct tag. The errors were in most cases names of persons that were tagged as organisations and organisations that received the category of ‘miscellanea’.

4.3 Evaluation of the Common Noun Semantic Tagger

The evaluation of the semantic tagging of common nouns is divided in its two main components: taxonomy induction (as explained in Section 3.3.2) and WSD (Section 3.3.3). For the first component, we evaluated the results of this task by random sampling 100 nouns from the taxonomy, in order to determine in which cases there was a correct ascending path to the top-node of the taxonomy. As was to be expected, a proportion of the nouns in the sample were polysemous, which means that we had to evaluate more hypernymy chains than nouns. In total, 144 chains were evaluated. Results are shown in Table 2.

Table 2: Figures of precision of hypernymy chains

Correct		Incorrect		Unfinished analyses	
n	%	n	%	n	%
91	63.20	30	20.83	23	15.97
Precision: 0.75 / Recall: 0.63					

Unfinished analyses are cases in which the chain production was interrupted for not reaching a semantic type after three recursive iterations, thus aborting the attempt. Errors in this task can be related to wrong detection of the hypernym in Wiktionary, but are mainly due to wrong sense disambiguation of the hypernym, which lead to a wrong connection of the noun with the ontology. There is of course ample room for improvement here that we leave for future work. There has been research on the use of supervised models for this task (e.g., Lopes et al., 2023; Tan et al., 2020), although the problem with such method arises when dealing with words not seen during the training phase.

Regarding WSD, the task of selecting the right sense of a polysemous noun in a sentence, we evaluated these models on a random sample of 100 sentences for 5 polysemous nouns (500 sentences in total). Table 3 shows the comparative figures of the performance of the different settings, including a random selection for general reference.

Table 3: Figures of precision of Spanish WSD results

Word	relex	fill-cooc	fill-cosine	target-cosine	random
<i>bodega</i> ‘cellar, winery’	0.75	0.33	0.50	0.59	0.67
<i>ratón</i> ‘mouse’	0.80	0.55	0.66	0.83	0.36
<i>operación</i> ‘operation’	0.47	0.23	0.19	0.25	0.21
<i>virus</i> ‘virus’	0.66	0.71	0.70	0.76	0.53
<i>medicina</i> ‘medicine, drug’	0.81	0.86	0.57	0.44	0.53
Mean	0.70	0.54	0.52	0.57	0.46

As observed in the table, Relex performs better than the other systems. Two observations can be made: 1) with values concentrated in the range .66 and .81 and with the lowest standard deviation (0.14), Relex seems relatively more stable in its results compared to the LLM-based methods. The others, with wider ranges (standard deviation greater than 0.2), seem more unpredictable; 2) showing better performance, Relex is also more parsimonious, being computationally simple and fast. With regards to error analysis, we observe that in most cases the root of the problem was the lack of sufficient data to make a correct prediction, i.e., many sentences do not contain enough lexical units to disambiguate, making it difficult even for humans to interpret the correct meaning. Consider, for instance, the case of *bodega*, which can mean ‘cellar’, ‘warehouse’ or ‘winery’, in a sentence like *Esta bodega contará con detectores de humo* ‘This warehouse will be fitted with smoke detectors’. A possibility to solve this problem in future work should be to consider larger context windows. Other problems are related with the metaphoric use of the words or with senses that are not attested in the dictionary.

4.4 Pattern Building Analysis

As a first evaluation of the quality of the produced patterns (see Section 3.4), we took a random sample of 5 verbs from the Verbario database, which, as explained in Section 2.2, contains manual analyses of corpus concordances necessary to extract the patterns. Table 4 shows the number of manually detected patterns per verb, the number of automatically detected patterns (considering the most frequent up to 7), the number of correct automatic patterns and the match between manual and automatic patterns.

Table 4: Evaluation of the patterns in a random sample of 5 verbs

Verb	Manual patterns	Automatic patterns	Correct	Match with manual
<i>amar</i>	2	3	3	1
<i>atrapar</i>	8	5	3	3
<i>beber</i>	8	7	6	1
<i>generar</i>	3	7	5	1
<i>iluminar</i>	5	2	2	2

A first impression upon examining these results is that, overall, the performance is quite acceptable, considering that it is a first attempt with this method and there are still many minor details to address. Part of the mismatch between the two sets is explained by the fact that the automatic patterns are only of the transitive type, while the manual annotation of the database is unrestricted and thus contains many more types of patterns, such as intransitive, ditransitive, pronominal and so on.

5. Conclusions and Future Work

In this paper, we presented a method for extracting CPA verb patterns from corpus, using a series of strategies to label sentences with syntactic and semantic information, in order to obtain the argument structure of the verb and the semantic types of each argument. The method shows a promising approach for a more effective pattern-based lexicography, as it provides the lexicographer with a set of patterns per verb automatically extracted, with acceptable precision. Pattern induction goes a step further from collocational analysis, which does not cover all the argument structure. We believe the method can be adapted to other languages with a reasonable volume of work.

At this stage, limitations of the method lay especially in the part of pattern building. We must deal with the splitting / lumping problem, trying to observe which level of the ontology is the most common for semantic typing, among other tasks to be addressed in the near future. More extensive evaluation is also necessary and, finally, we should test the system with a team of lexicographers using these results in their job.

References

- Agirre, E., de Lacalle, O. L., & Soroa, A. (2014). Random walks for knowledge-based word sense disambiguation. *Computational Linguistics*, 40(1), 57–84.
- Akbik, A., Bergmann, T., & Vollgraf, R. (2019b). Multilingual sequence labeling with one model. In *NL DL 2019, Northern Lights Deep Learning Workshop*.
- Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., & Vollgraf, R. (2019a). FLAIR: An Easy-to-use framework for state-of-the-art NLP. In J. Burstein, Ch. Doran, & Th. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)* (pp. 54–59). Association for Computational Linguistics.

Amsler, R. (1981). A taxonomy for English nouns and verbs. In *Proceedings of 19th Annual Meeting on ACL (Morristown, NJ, USA)*, 133–138.

Baisa, V., Bradbury, J., Cinková, S., El Maarouf, I., Kilgarrieff, A., & Popescu, O. (2015). SemEval-2015 Task 15: A CPA dictionary-entry-building task. In P. Nakov, T. Zesch, D. Cer, & D. Jurgens (Eds.), *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)* (pp. 315–324). Association for Computational Linguistics.

Battaner, P., & Renau, I. (2011). El proyecto DAELE verbos: un diccionario de aprendizaje de español como lengua extranjera. *II Encuentros (ELE) Comillas: el profesor de ELE: metodología, técnicas y recursos para el aula*.

Bevilacqua, M., Pasini, T., Raganato, A., & Navigli, R. (2021). Recent trends in word sense disambiguation: A survey. In Z.-H. Zhou (Ed.), *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21)* (pp. 4330–4338). International Joint Conferences on Artificial Intelligence.

Chodorow, M., Byrd, R., & Heidorn, G. (1985). Extracting semantic hierarchies from a large on-line dictionary. In *Proceedings of the 23rd annual meeting on ACL*, 299–304.

Colman, L., & Tiberius, C. (2018). A good match: A Dutch collocation, idiom and pattern dictionary combined. In J. Čibej, V. Gorjanc, I. Kosem, & S. Krek (Eds.), *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts* (pp. 233–246). Ljubljana University Press.

CROATPAS: <https://croatpas.baisa.cz/> [Accessed: 25/7/24]

DiMuccio-Failla, P. V., & Giacomini, L. (2017). Designing a learner's dictionary based on Sinclair's lexical units by means of Corpus Pattern Analysis and the Sketch Engine. In I. Kosem, C. Tiberius, M. Jakubiček, J. Kallas, S. Krek, & V. Baisa (Eds.), *Electronic lexicography in the 21st century. Proceedings of eLex 2017 Conference* (pp. 437–457). Lexical Computing.

DiMuccio-Failla, P. V., & Giacomini, L. (2022). A proposed microstructure for a new kind of active learner's dictionary. *Lexicographica*, 38(1), 475–499.

Giacomini, L., & DiMuccio-Failla, P. (2019). Investigating semi-automatic procedures in pattern-based lexicography. In I. Kosem, T. Zingano Kuhn, M. Correia, J. P. Ferreira, M. Janssen, I. Pereira, J. Kallas, M. Jakubiček, S. Krek, & C. Tiberius (Eds.), *Electronic lexicography in the 21st century: Smart lexicography. Proceedings of the eLex 2019 Conference* (pp. 490–505). Lexical Computing.

Gutiérrez Fandiño, A., Armengol Estapé, J., Pàmies Massip, M., Llop Palao, J., Silveira Ocampo, J., Carrino, C. P., Armentano Oller, C., Rodríguez Penagos, C., González Agirre, A., & Villegas, M. (2022). MarIA: Spanish Language Models. *Procesamiento del Lenguaje Natural*, (68), 39–60.

Grefenstette, G., & Hanks, P. (2023). Competing views of word meaning: word embeddings and word senses. *International Journal of Lexicography*, 36(2), 211–219.

Guthrie, L., Slator, B., Wilks, Y., & Bruce, R. (1990). Is there content in empty heads? In H. Karlgren (Ed.), *Proceedings of the 13th International Conference on Computational Linguistics, COLING'90* (pp. 138–143). Association for Computational Linguistics.

Hanks, P. (2004a). Corpus Pattern Analysis. In G. Williams, & S. Vessier (Eds.), *Proceedings of the 11th Euralex International Congress* (Vol. 1, pp. 87–97). Université de Bretagne-Sud.

Hanks, P. (2004b). The syntagmatics of metaphor and idiom. *International Journal of Lexicography*, 17(3), 245–274.

Hanks, P. (2013). *Lexical Analysis: Norms and Exploitations*. The MIT Press.

Hanks, P., & Ježek, E. (2008). Shimmering lexical sets. In E. Bernal, & J. DeCesaris (Eds.), *Proceedings of the 13th EURALEX International Congress* (pp. 391–402). Universitat Pompeu Fabra.

Hanks, P., & Pustejovsky, J. (2005). A pattern dictionary for natural language processing. *Revue Française de Linguistique Appliquée*, 10(2), 63–82.

Huang, L., Sun, C., Qiu, X., & Huang, X. (2019). GlossBERT: BERT for word sense disambiguation with gloss knowledge. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)* (pp. 3509–3514). Association for Computational Linguistics.

Ježek, E., & Hanks, P. (2010). What lexical sets tell us about conceptual categories. *Lexis. Journal in English Lexicology*, 4, 7–22.

Ježek, E., Magnini, B., Feltracco, A., Bianchini, A., & Popescu, O. (2014). T-PAS; A resource of Typed Predicate Argument Structures for linguistic analysis and semantic processing. In N. Calzolari et al. (Eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* (pp. 890–895). ELRA.

Kilgariff, A., & Renau, I. (2013). EsTenTen, a vast web corpus of Peninsular and American Spanish. *Procedia-Social and Behavioral Sciences*, 95, 12–19.

Kilgariff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, 1, 7–36.

Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In V. DeBuys (Ed.), *Proceedings of the 5th Annual International Conference on Systems Documentation (SIGDOC'86)* (pp. 24–26). Association for Computing Machinery.

Lopes, A., Carbonera, J., Schmidt, D., Garcia, L., Rodrigues, F., & M. Abel (2023). Using terms and informal definitions to classify domain entities into top-level ontology concepts: An approach based on language models. *Knowledge-Based Systems*, 265, 110385.

Marini, C. (2022). CroaTPAS: A survey-based evaluation. In H. Bunt (Ed.), *Proceedings of the 18th Joint ACL - ISO Workshop on Interoperable Semantic Annotation within LREC2022* (pp. 76–80). ELRA.

Marini, C., & Ježek, E. (2019). CROATPAS: A resource of corpus-derived typed predicate argument structures for Croatian. In R. Bernardi, R. Navigli, & G. Semeraro (Eds.), *Proceedings of the Sixth Italian Conference on Computational Linguistics*.

Nazar, R., & Renau, I. (2016). A taxonomy of Spanish nouns, a statistical algorithm to generate it and its implementation in open source code. In N. Calzolari et al. (Eds.), *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16)* (pp. 1485–1492). ELRA.

PDEV: <https://pdev.org.uk/> [Accessed: 25/7/2024]

Pustejovsky, J., Hanks, P., & Rumshisky, A. (2004). Automated induction of sense in context. *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, 924–930.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. (2020). Stanza: A Python natural language processing toolkit for many human languages. In A. Celikyilmaz, & T.-H. Wen (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 101–108). ACL.

Renau, I. (2012). *Gramática y diccionario: las construcciones con se en las entradas verbales del diccionario de español como lengua extranjera* [Doctoral dissertation, Universitat Pompeu Fabra].

Renau, I., Nazar, R., Castro, A., López, B., & Obreque, J. (2019). Verbo y contexto de uso: un análisis basado en corpus con métodos cualitativos y cuantitativos. *Revista Signos*, 52(101), 878–901.

Sinclair, J. M. (1991). *Corpus, Concordance, Collocation*. Oxford University Press.

Sinclair, J. M. (2004). *Trust the Text, Language, Corpus and Discourse*. Routledge.

Straka, M. (2018). UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In D. Zeman, & J. Hajič (Eds.), *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies* (pp. 197–207). ACL.

Tan, Y., Wang, X., & Jia, T. (2020). From syntactic structure to semantic relationship: hypernym extraction from definitions by recurrent neural networks using the part of speech information. In J. Z. Pan et al. (Eds.), *The Semantic Web – ISWC 2020. 19th International Semantic Web Conference* (pp. 529–546). Springer.

Tiedemann, J. (2009). News from OPUS — A collection of multilingual parallel corpora with tools and interfaces. In N. Nicolov, K. Bontcheva, G. Angelova, & R. Mitkov (Eds.), *Recent Advances in Natural Language Processing: Vol. V* (pp. 237–248). John Benjamins.

T-PAS: <https://tpas.fbk.eu/> [Accessed: 25/7/24]

Verbario: <http://www.verbario.com> [Accessed: 25/7/24]

Wolf, T. et al. (2020). Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 38–45). ACL.

Wiktionary: <https://en.wiktionary.org/wiki/Spanish> [Accessed: 25/7/24]

Woordcombinaties: <https://woordcombinaties.ivdnt.org/> [Accessed: 25/7/24]

Acknowledgements

We would like to express our gratitude to the evaluators for their valuable comments. This research received funds from project Fondecyt Regular nr 1231594 and ESMAS-ES+ (PID2022-137170OB-I00) funded by MCIN/AEI//FEDER “Una manera de hacer Europa”. We want to thank Patrick Hanks for all these wonderful years of learning, inspiration and guidance. You will always be in our hearts!

Contact information

Irene Renau

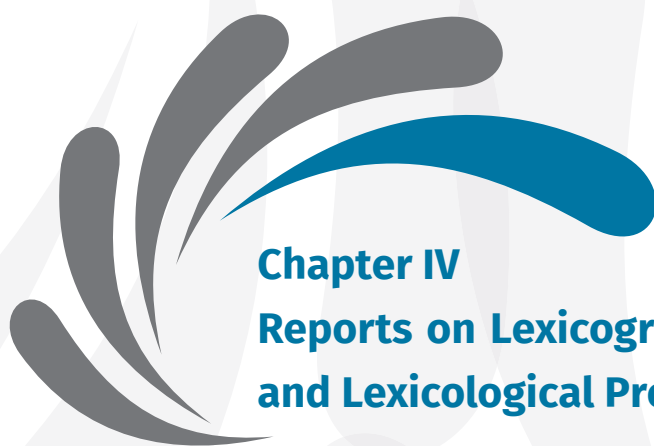
Instituto de Literatura y Ciencias del Lenguaje, Pontificia Universidad Católica de Valparaíso
irene.renau@pucv.cl

Rogelio Nazar

Instituto de Literatura y Ciencias del Lenguaje, Pontificia Universidad Católica de Valparaíso
rogelio.nazar@pucv.cl

Daniel Mora

Instituto de Literatura y Ciencias del Lenguaje, Pontificia Universidad Católica de Valparaíso
daniel.mm91@gmail.com



Chapter IV

Reports on Lexicographical and Lexicological Projects

Juris Baldunčiks and Silga Sviķe

PAGES OF LATVIAN HISTORICAL SLANG DICTIONARY: *DZERŠANA* ('DRINKING')

Abstract For Latvian linguists, the study of slang was not a topical matter until 1970. The literary language and dialects have always been perceived as research priority, and the non-literary language was not considered an object of serious scientific work for a long time. There was a more or less pronounced derogation of the non-literary language. Only a few enthusiasts showed scientific interest in it. Recently, research on Latvian slang has taken major steps with the publication of a dictionary *Latviešu valodas slenga vārdnīca* (Latvian Slang Dictionary) by Ojārs Bušs and Vineta Ernstone in 2006. This study aims to describe the challenges and solutions that have arisen during the development of the unpublished *Latviešu vēsturiskā slenga vārdnīca: dzeršana* (Latvian Historical Slang Dictionary: *dzeršana* ('drinking')). The analysed linguistic material is compiled from written sources (from the 17th century onwards), speech notes (from the late 1970s onwards), and student surveys (from the second half of the 1990s onwards). Since *Latviešu valodas slenga vārdnīca* contains mainly lexis from the last 30–60 years, the paper basically focuses on the period from the origins of Latvian slang until the Second World War.

Keywords Latvian historical slang; special slang; slang dictionary; *dzeršana*

1. Introduction

There are several reasons for the relatively late start of research into Latvian slang. Firstly, the study of dialects and literary language has always been valued and considered more important than other layers of language. Secondly, especially during Soviet times, the non-literary language was associated with an official negative attitude and even a generally undifferentiated battle against this part of the language. Probably, thanks to the studies of English slang in Soviet linguistics and efforts in other neighbouring countries, some studies on slang appeared (e.g., Bušs, 1979). It took several years for the ice to finally break in this field. Although there is a dictionary of slang in Latvian (*Latviešu valodas slenga vārdnīca*) by Ojārs Bušs and Vineta Ernstone (2006), some problems directly or indirectly related to slang have not yet received sufficient attention. One of the unexplored topics is the historical development and research of Latvian slang from its origins. The aim of this study is to provide an insight into the Latvian historical slang units that characterise the thematic field – 'drinking' (in Latvian – *dzeršana*). To achieve this goal, a descriptive approach is used and an analysis of an unpublished dictionary of Latvian historical slang in the thematic field – drinking – developed by Juris Baldunčiks (1950–2022) is carried out. This slang dictionary has not yet been completed, but is in the development phase. A brief description of the dictionary article is given in Chapter 4 of this paper.

2. The Need for Research on Latvian Slang

The Cambridge Dictionary defines slang as a very informal language used by particular groups of people; it is more often spoken than written and includes words that are not polite. So, these words, which belong to the level of slang, cannot be considered literary ones. The use of the term ‘slang’ in Latvian – *slengs* – was pioneered by Bušs in 1979, who used it to refer to all non-dialectal, non-literary vocabulary of the Latvian language (Bušs, 1979, pp. 68–77). In the following twenty to thirty years, the Latvian term *slengs* (slang) was also used in researches on the language of young people, and Ernstone’s (1998, pp. 32–36; 2000, pp. 23–24) studies of non-literary vocabulary also included this term. Bušs (1979) divides slang into a common and special one, where a common slang consists of the common non-literal non-dialectal elements of language, and a special slang consists of the non-literal non-dialectal lexical layers of professionally or socially differentiated groups. This study analyses special slang because it focuses on a specific thematic group – ‘drinking’ (Latvian – *dzeršana*) and looks specifically at historical slang – from the beginning of the collection of slang items to the Second World War.

Firstly, there is a need for research that would shed some light on the historical development of Latvian slang. Here it is also necessary to return to the question of the differentiation of the Latvian language into sub-systems, which has so far been treated incompletely, sometimes even with a biased ideological overlay. Sufficiently rich linguistic material needs to be gathered to be able to pinpoint the historical period when the stylistic understanding of non-literal elements of language began to form and consolidate in the consciousness of educated Latvians (this does not mean vulgarisms and the explicitly “naughty” vocabulary against which pastors already spoke in the Old Latvian period).

Secondly, there is a need to correct some of the positions in contemporary Latvian stylistics, as the conservative views of Rozenbergs have dominated this field for too long (his work in the last decades in the study of non-literary language was basically ignored). Stylistics is a subfield of linguistics in which subjective judgements are much more common than in other fields, so a system of lexical analysis that minimises subjectivism is needed. For example, Rozenbergs (2000, p. 129) says that ‘trend words’ (Latvian – *modes vārdi*) can be at different levels in relation to the norms of literary language:

1. they can be at the level of literary written language (e.g., Latvian – *cīnīties* (‘fight’), *joma* (‘sphere’), *klāsts* (‘range’), *lēģenda* (‘legend’), *viennozīmīgi* (‘straightforward’));
2. they may be at the level of literary written language (e.g., Latvian – *foršs* (‘cool’ as slang for ‘very good’), *senči* (‘rents’ as slang for ‘parents’), *stilīgs* (‘vogueish’ as slang for ‘stylish’), *perfekti* (‘ace’ as slang for ‘good, correct, intelligent’));
3. they may be at low colloquial level (*učuks* (‘teach’ as slang for ‘teacher’), *vot* (‘jigs’ as slang for ‘so there’), *kruts* (‘zooly’ as slang for ‘cool’)).

It is not clear what criteria the linguist used to form the second and third groups of examples. One doesn't even have to be a stylistic expert to realise that only Latvian words *stilīgs* and *perfekti* fit into everyday speech. It is also questionable whether the words of the third group really illustrate a low colloquial level. All the words in this group, as well as *foršs* and *senči*, are included in *Latviešu valodas slenga vārdnīca*; this leads to reflections on the stylistic division of lexis used in Latvian stylistics.

3. Difficulties and Possible Solutions in Latvian Slang Research

It would be necessary to analyse in detail a wide range of border lexis from the literary and non-literary layers, as well as non-literary lexis, and to develop a new stylistic classification. The task also involves defining the boundaries of slang. This is one of the most difficult questions in slang research. Being aware of the diversity of this linguistic layer, it is obviously necessary to look for features that allow linguistic units to be included into or excluded from slang. For example, the stylistic division of vocabulary is supposed to impose an order on those lexical layers that do not belong to the neutral lexicon, but the experience of some dictionaries makes this doubtful, since a single restrictive reference, e.g., *sarunvaloda* ('colloquial speech') can denote linguistic units with different stylistic intensity, *smuks* ('pretty'), *stripa* ('line') and *rukāt* ('swot'), *šļuka* ('shot') in *Latviešu literārās valodas vārdnīca* (LLVV; Dictionary of Latvian Literary Language); the last two words are also included in the *Latviešu valodas slenga vārdnīca* (LVSV).

It is possible to use a stylistic shade scale (Skrebnev, 2000, p. 56), in which the neutral base lexis forms a relative zero bar and lexical items with plus and minus signs are arranged above and below it. There is a gradation in both directions, so there are also levels for lexis with a minus sign:

1. minimally coloured colloquial vocabulary, which in most cases is not at all perceived by the language users and is basically used without situational restrictions,
2. moderately coloured colloquial vocabulary, which most Latvian speakers can distinguish from neutral vocabulary without much effort, and
3. strongly coloured non-literal vocabulary, which the average statistical Latvian speaker uses only in certain situations and with certain communicants (an exception might be the communication environment of today's young people).

We can assume that slang includes all lexical items that are below the minimally marked colloquial lexicon. This is roughly the approach taken by the authors of *Latviešu valodas slenga vārdnīca* (Bušs & Ernstone, 2006, p. 7). The same can be said about many English slang dictionaries. Of course, this scale can also be subjective, based on various factors (knowledge, upbringing and life experience, breadth of social contact, etc.).

It is necessary to study various sources of Latvian slang and their historical changes in more detail. There are many borrowings in Latvian slang. They are

often referred to as barbarisms, but the usual 21st century interpretation of this term no longer seems scientific enough, as it is largely based on insufficiently objective criteria. The division of barbarisms (initial and developmental barbarisms) by Ozols (1961; 1967, pp. 452–460) is also the view of a linguistic culture specialist who has not himself studied the mechanism of acquisition of words of foreign origin, so the reasoning of his classification does not really coincide with the actual motives and circumstances of word borrowing, as well as the prevailing ideas in research on borrowings. Perhaps this concept, specifically used in Latvian linguistics, should be gradually excluded from the study of borrowed and non-literary languages.

4. Description and Characteristics of the Lexis of the Latvian Historical Slang Dictionary

In order to find out the origins of Latvian slang in the thematic field of ‘drinking’, first it is necessary to summarise the lexis related to alcoholic beverages and their consumption found in Latvian written sources, to get acquainted with the development of the corresponding lexis in contact languages, as well as to look into the history of the production and distribution of alcoholic beverages.

The most important drinks in this group are beer, wine, and vodka (including other spirits). Wine is a very old drink, dating back thousands of years. In Latvian, the word *vīns* (‘wine’) is an ancient derivative (found in 16th century written sources), most probably from Germanic languages, where it was used as early as the 8th century (Kluge, 1999, p. 882).

Beer is also a very old drink (Dumpe, 2001). The Latvian word *alus* (‘beer’) is probably an old Germanic derivative, but other opinions have been expressed about the etymology of the word (Karulis, 2001, p. 68).

As early as in the 17th century, Latvian written sources contain both the word *vīns* (‘wine’) and the word *alus* (‘beer’) in various lexical combinations. However, beer seems to have been more popular, since it was with the word *alus* that appears in written sources as one of the first slangisms: *alus muša* (‘hard drinker’), in Latvian – ‘*dzērājs*’ (Fürecker, comp. with German *Bierfliege*).

Primitive distillation of alcohol was known in some parts of Europe probably as early as in the 11th century. In the 12th century, spirits were made in both southern and western Europe (Americana, 1992), in the 15th century, the production of different types of vodka expanded, and in the 16th century spirits spread throughout Europe. As early as in the 13th century, the Latin (*aqua vitae*, *aqua ardens*, etc.) also appeared in Middle High German *brantwīn* (from the German verb *brennen* with the meaning of ‘distilling’) (Kluge, 1999, p. 131). By the early 17th century, the word had been adopted in several Germanic languages (e.g., Low German, Dutch, English) and some other languages. In Latvian, it was first recorded in G. Mancelius’ dictionary.

The linguistic material collected by Baldunčiks (more than 600 entry words and phrases) has been compiled from written sources (from the 17th century onwards), speech recordings (from the late 1970s onwards), and student surveys (from the second half of the 1990s onwards). There are basically only three types of sources in the group of articles: 1) dictionaries, 2) fiction, 3) periodicals (mainly humorous essays, feuilletons, short reports on incidents related to drinking, legal or illegal sale of alcoholic beverages, etc., humorous dialogues). Since *Latviešu valodas slenga vārdnīca* published in 2006 largely contains lexis from the last 30–60 years, this article mainly focuses on the period from the beginning of Latvian slang to the Second World War.

Slang terms related to drinking (drinking to get drunk) basically form four groups:

1. terms for drinks by type (*blakšūdens* (for ‘whisky’), *dzimtene* (for ‘vodka’), *konis* (for ‘cognac’), *šampis* (for ‘champagne’) etc.) or type and volume (*ampula* (for ‘a very small bottle’), *kočiņš* (for ‘a vessel of a capacity of one quarter of a litre’), *polšs* (for ‘a half litre bottle of vodka’) etc.);
2. drinking activity or process designations (*iekampt*, *ķemmēt*, *mest*, *plaut* (all Latvian slangisms for ‘to drink’), etc.);
3. designations of drunkenness (*autā*, *labā dūšā*, *pilnā*, *šarabī* (all Latvian slangisms for ‘drunk’), etc.);
4. designations of drinkers (*alkašs*, *kodējs*, *metējs*, *siņķa* (all Latvian slangisms for ‘a drunkard’), etc.)

In addition, there are other smaller subgroups, such as designations of drinking, alcohol outlets, alcohol containers, and hangovers.

As the dictionary is in the development phase, below is a small insight into the key parts of the dictionary article with explanatory notes. The current structure of the dictionary is a list of entry words (slang words and other slang language units), for which explanations in Latvian and examples of usage from various historical sources are searched and added. One example of a dictionary article is shown in Example 1 below. The example shows how *baltais* // *baltiņš* // *baltais dzidrais* – that are slang words for Latvian *degvīns* (‘vodka’) – were used in the source, the Latvian newspaper *Cīņa*.

(1) Example of a Latvian historical slang dictionary article for designations of *degvīns* (‘vodka’)

baltais // *baltiņš* // *baltais dzidrais* – *degvīns*

[..] ceturtajā vietā izstāvēju *balto*. (Cīņa No. 202, 1877, p. 2)

[..] jānopelna vēl viens *baltiņš*. (Cīņa No. 202, 1877, p. 2)

baltais dzidrais (Cīņa No. 116, 1954, p. 4)

The example shows that the structure of the dictionary article is simple, basically offering an explanation of slangisms in Latvian literary language and examples of usage from historical sources. Phraseologisms, phrasemes and idioms (e. g. *apdzēries kā lūks* (for ‘very drunk’), *āžu dīrāšana* (for ‘vomiting’)) are also included as entry words and they have the same structure.

5. Conclusions

For a long period of time, research on Latvian slang lexis remained on the periphery of interest, as other theoretical problems had to be solved. Slang, as an unexplored part of the Latvian language, is valuable for researching because it shows a specific language layer spoken by a certain group of people under certain circumstances. Research into specialised slang materials requires a thorough collection of data. Historical slang about the thematic group ‘drinking’ reflects a specific set of linguistic units, which can be divided into groups and compiled into a dictionary, including examples and explanations of slangisms. This resource could be useful for researchers and translators in understanding the language, society, and culture of the time.

References

- Americana (1992). *The Americana Annual 1992*. An Encyclopedia of the Events of 1991. Yearbook of the Encyclopedia Americana. Grolier incorporated.
- Bušs, O. (1979). Latviešu valodas nedialektālā neliterārā leksika jeb slengs. *LPSR Zinātņu akadēmijas Vēstis*, 68–77.
- Bušs, O., & Ernstsone, V. (2006). Priekšvārds. *Latviešu valodas slenga vārdnīca* (pp. 6–9). Norden AB.
- Cambridge Dictionary (n.d.). Slang. In *Dictionary.Cambridge.org*. Retrieved May 18, 2024, from <https://dictionary.cambridge.org/dictionary/english/slang>
- Cīņa* No. 202 (1877). [Newspaper] (p. 2).
- Dumpe, L. (2001). *Alus tradīcijas Latvijā*. Latvijas Vēstures institūta apgāds.
- Ernstone, V. (1998). Slenga elementi mūsdienu reklāmā. *Vārds un tā pētīšanas aspekti: rakstu krājums*, 2. (pp. 32–36). Liepājas Pedagoģijas Akadēmija.
- Ernstsone, V. (2000). Slengismi mūsdienu latvieša runā, to situatīvais aspekts. *Zinātnes Vēstnesis*. 27. 10. 2000. (pp. 23–24).
- Fürecker, Ch. (1685 [1997]). Lettisches und Teutsches Wörterbuch zusahmen geschrieben und mit fleiß gesamlet Von Christopher Fürecker. I. T. G. Fennell. *Fürecker's dictionary: the first manuscript*. Latvijas Akadēmiskā bibliotēka.
- Karulis, K. (2001). *Latviešu etimoloģijas vārdnīca*. Avots.

Kluge, F. (1999). *Etymologisches Wörterbuch der deutschen Sprache*. 23. erw. Aufl. Bearb. von E. Seebold. Walter de Gruyter.

LVSĻ = *Latviešu valodas slenga vārdnīca*. Bušs, O., & Ernstone, V. (2006). *Latviešu valodas slenga vārdnīca*. V. Zvaigzne ABC.

LLVV = *Latviešu literārās valodas vārdnīca*. (1972–1996), 1–8. Zinātne.

Ozols, A. (1961). *Latviešu tautasdziesmu valoda*. Latvijas Valsts izdevniecība.

Ozols, A. (1967). Latviešu rakstu valodas un literārās valodas vēstures pētīšanas metodes un sistēma. *Raksti valodniecībā* (pp. 452–460). Zinātne.

Rozenbergs, J. (2000). Modes vārdi. *Linguistica Lettica*, 7, 128–133.

Skrebnev, Y. (2000). Osnovy stilistiki anglijskogo yazyka [The basics of the stylistics of the English language]. Astrel' (in Russian) [Скребнев, Ю. (2000). Основы стилистики английского языка. Астрель].

Acknowledgements

This research has been funded by the State Culture Capital Foundation, agreement No. 2024-1-LIT047. Thanks to Ingrida Irbe, widow of Professor Juris Baldunčiks, for providing access to the materials collected by her husband for this research. Thanks to Natalja Malašonoka for improving the English language of the manuscript.

Contact information

Juris Baldunčiks

Ventspils University of Applied Sciences

Silga Sviķe

Ventspils University of Applied Sciences

silga.svike@gmail.com

María Auxiliadora Barrios Rodríguez

DIRETES, A SPANISH MONOLINGUAL DICTIONARY BASED ON LEXICAL-SEMANTIC RELATIONS

Abstract *Diretes* is a Spanish monolingual e-dictionary based on Lexical-Semantic Relations which are formalized by Lexical Functions, a formal tool explored within the Meaning-Text Theory. This dictionary consists of a relational database which aims to reflect the cognitive links of the lexicon through a network of semantic and lexical associations. Currently it contains more than 100,000 collocations and semantic relations, and includes dictionary entries for simple words, compound words, idioms and formulaic expressions. The majority of the definitions of each entry consist of two parts: a minimal definition (written with basic words) and an expanded definition (which collects world knowledge data). The structure of the entries of the dictionary for formulaic expressions, however, are more complex as they also include a pragmatic function, a prototypical scenario, and certain other specific information. Semantic relations include synonymy, antonymy, hyponymy and other productive relations not utilized to date in other dictionaries. Lexical Functions allow the user of the dictionary to consult the entries, including idioms, through both onomasiological and semasiological approaches. This paper summarizes the purpose of the creation of *Diretes* and the idiosyncracies of this dictionary.

Keywords relational database; lexical functions; collocations; semantic relations; idioms; formulaic expressions

1. Why a New Spanish Monolingual Dictionary?

Spanish Lexicography is an ancient discipline with its inception in 1492 via the first *Spanish- Latin Lexicon* of Antonio Nebrija. For a century, all the published dictionaries in Spain were bilingual, amongst them the *Spanish-Arabic Lexicon* of Pedro Alcalá (1505), the *Spanish-French Lexicon* of Nebrija (1516), and the *Spanish-Italian Lexicon* of Casas (1570).¹ Particularly relevant were the bilingual Lexicons developed by the Spanish missionary in America, some of them under the influence of the book of grammar and the dictionary of Nebrija, such as the *Lexicon of the Mexican Language* of Andrés de Olmedo (1547) or that from de Tzendal and other languages, de Francisco de Cepeda (1560).

The first two Spanish monolingual dictionaries were published in 1611 by Covarrubias and Rosal. Throughout the seventeenth century several bilingual dictionaries were published with no other Spanish monolingual dictionary coming into existence until 1729 when the Spanish Royal Academy published the first of twenty eight different versions of its Spanish monolingual dictionary. After almost three centuries it can be asserted without doubt that this dictionary has become the most popular dictionary

¹ All the dictionaries mentioned in this paragraph can be referenced in the academic website <https://apps.rae.es/nttle/SrvltGUISalirNttle>.

amongst Spanish speakers both in Spain and abroad. Its influence was such that over two centuries all of the Spanish monolingual dictionaries were a slightly modified copy of the last version of this academic dictionary.

Over the last century and a half, however, some notable monolingual dictionaries have been published by lexicographers with their own particular layout and original definitions: the *DCR* by Cuervo, which is focused on grammatical features;² the *DCECH* by Corominas and Pascual, which is a historical dictionary;³ the *DUE* by María Moliner, which was the first standard dictionary containing some collocations;⁴ the *DEA* by Seco et al., the authors of which applied the criteria of substitution when writing the definition;⁵ and the two combinatorial dictionaries *Redes* and *Práctico*, by Bosque (2004; 2006), focused on collocations, which do not contain definitions but meanings based on semantic relations: their content and approach see them unique in the current landscape of international lexicography.⁶

Despite the five aforementioned dictionaries breaking new ground there are some potential lexicographic innovations that have not yet been applied: neither Spanish nor other languages dictionaries have yet to be completed following the Meaning-Text Theory criteria. This linguistic model has been viewed as the “only really new thing that has appeared in recent dictionaries” (Simone, 2016, p. 29). The Spanish lexicographic tradition of the two previously mentioned centuries formed the basis for these innovative dictionaries of the last century and a half. By the same token, we take into consideration the fact that the richness of the unique Spanish combinatorial dictionaries, *Redes* and *Práctico*, has significantly facilitated the creation of *Diretes*. The reason being our dictionary is based on data derived from various sources, the most important being the collocations and semantic structures present in both dictionaries.⁷ Due to their great depth, *Diretes* aspires to be the first monolingual Spanish dictionary to be completed within the MTT model, although, as will be shown, we have added some particularities which are not present in canonical dictionaries of the model.

The next two sub-sections concentrate on demonstrating the relevance of this model when compiling dictionaries (§ 1.1), and particularly why a dictionary such as *Diretes* would be necessary (§ 1.2). Section 2 concentrates on the definitions in *Diretes*. Section 3 presents the methodology followed when working with semantic relations. Section 4 summarizes the methodology for the formalization of the collocations. Section 5 portrays the novelties around the idioms. And finally, Section 6 presents the conclusions.

² The *Diccionario de construcción y régimen de la lengua castellana* by Rufino Cuervo (1886), <https://www.rae.es/archivo-digital/diccionario-de-construccion-y-regimen-de-la-lengua-castellana>.

³ The *Diccionario crítico etimológico castellano e hispánico* by Joan Corominas and José A. Pascual (1980).

⁴ The *Diccionario de Uso del Español* by María Moliner (1966).

⁵ The *Diccionario del Español Actual*, by Manuel Seco, Olimpia Andrés, Gabino Ramos (1999).

⁶ Some other non-academic lexicography has been developed over the last thirty years (for further information, see Bosque & Barrios, 2017).

⁷ Lack of space prevents us from mentioning all the sources we use. Most of them are open access corpora rich in collocational information, such as the esTenTen, a Spanish corpus from the Sketch Engine, <https://www.sketchengine.eu/estenten-spanish-corpus/>.

1.1 Dictionaries Within the Meaning-Text Theory

The Meaning-Text Theory (hereinafter MTT) is a lexicological and lexicographic model which proposes a systematic methodology for dictionaries based on formal tools referred to as Lexical Functions (Mel'čuk 1984; 1996; 2014; 2023; Mel'čuk et al, 1995; 2007; 2018; Apresjan 1992; 2009; Polguère 2000; 2003; 2007). Lexical Functions (LFs) facilitate the classification of lexical relations by means of the formalization not only of collocations but also of paradigmatic relations, such as synonymy, antonymy, and semantic derivatives. For instance, within the MTT the semantic relation between *grain* and *wheat* is similar to that between *drop* and *water*: both are associated to one LF called Sing, which means '(the minimal) part of'. This LF allows us to formalize both collocations as shown in (1). Similarly used is the relation between *flock* and *sheep*, and *string* and *garlic*, both of them related to an LF called Mult, which means 'group of', as shown in (2):

1. Sing(wheat) = grain (of); Sing(water) = drop (of)
2. Mult(sheep) = flock (of); Mult(garlic) = string (of)

The first MTT dictionary was a monolingual French print dictionary (Mel'čuk, 1984), which was followed by the *Lexique actif du français* (Mel'čuk & Polguère, 2007), a printed Lexicon containing around 20,000 semantic relations and collocations. The leap from the printed dictionary to the on-line dictionary for the French language was made with the *Réseau Lexical du Français* (Polguère, 2014), accessible via website (<https://spiderlex.atilf.fr/fr>), which was based on the French dictionary *Dicouèbe* (<http://olst.ling.umontreal.ca/dicouebe/index.php>).

The Spanish language relies on the *DICE*, an on-line dictionary, the vocabulary of which was focused on nouns of emotions in its first version (Alonso Ramos, 2004). The last version contains collocations, with its entries being nouns from all lexical fields (<http://www.dicesp.com/paginas>).

There is also an automatic Russian-English translator (<http://cl.iitp.ru/etap3>) based on LFs and some dictionaries for Russian. Although most of the dictionaries developed within the MTT (as the recently appointed) make use of the LFs, there are a few exceptions. A case in point *Russian Synonyms Dictionary* which was directed by Jury Apresjan (2018) as well as a monolingual Russian dictionary, *The Active Dictionary of Russian* (an ongoing project of the Russian Language Institute also directed by Apresjan). The last one includes the definitions of all senses of words, the government patterns, the constructions in which the word is used, and certain other linguistic information.⁸ The definitions are "systemic, comprehensive, non-tautological, and explanatory of co-occurrence properties of lexical items" (Apresjan, 2011).

Particularly prolific are the applications of the MTT to terminology (see, amongst others, L'Homme, 2012; 2018; L'Homme et al, 2016), such as the *Dictionnaire fondamental de l'informatique et de l'Internet* (https://olst.ling.umontreal.ca/?page_id=53), a legal

⁸ I would like to express my gratitude to Igor Boguslavsky for the oral information about this dictionary.

terminology database (https://olst.ling.umontreal.ca/?page_id=1357), the *Dictionnaire fondamental de l'environnement* (https://olst.ling.umontreal.ca/?page_id=214); and two projects that apply the MTT and the semantic frames to the terminology: the *Framed Dicoenviro* (https://olst.ling.umontreal.ca/?page_id=2364), and the *Framed DicoInfo* (https://olst.ling.umontreal.ca/?page_id=3243).

1.2 Innovations in Diretes

Diretes is a Spanish monolingual e-dictionary based on Lexical-Semantic Relations which are formalized by LFs. The name of the dictionary is an acronym of *Dlccionario RETicular del ESpañol* (Spanish Reticular Dictionary) which aims to express the deliberate intention of reflecting the cognitive links of the lexicon through a network of semantic and lexical associations (Polguère, 2014; Barrios, 2020). As with most of the modern dictionaries, *Diretes* is a relational database (Atkins & Rundell, 2008). Currently it contains more than 100,000 collocations and semantic relations (Barrios, 2022a). Lexical inheritance (Mel'čuk & Wanner, 1996; Barrios, 2010) provided the base of a third part of the collocations, meaning that the formalization of more than 25,000 collocations in the database was performed automatically (as we will see in Section 4). A pilot test of *Diretes* is available at <http://diretes.es/>.

The main particularities of the dictionary, amongst others, are the following:

- a. *Diretes* includes dictionary entries not only for words but also for idioms, compound words and formulaic expressions.
- b. Definitions of words, idioms and compound words consist of two parts: a minimal definition (written in basic words) and an expanded definition (which collects world knowledge data). This structure allows us to offer concomitantly definitions corresponding both to monolingual dictionaries of Spanish and to encyclopedic dictionaries. In fact, minimal definitions are linguistic definitions based on semantics whilst expanded definitions are based on data from knowledge of the world (Roche, 2015). Both definitions provide ontological features because the minimal definition relies on ontological relations by means of the semantic label, whilst the expanded one relies on them by means of the expanded information. The latter is particularly useful for Terminology, Second Language Teaching and other applied fields of linguistics (Frawley, 1997).⁹
- c. The structure of the entries of the dictionary for formulaic expressions are more complex as they also include a pragmatic function (Barrios, 2022b), a prototypical scenario (Blanco, 2013), and certain other information. For instance, for the Spanish expression *¡Pide un deseo!* (lit. *Make a wish!*) the entry presents the following data: a) the pragmatic function (to generate empathy); b) typical scenarios (such as after blowing out the candles on the cake); c) the attitude of the speaker (more a childish game than of a

⁹ In *Diretes* we have yet to work with terminology, with the exception being for some words from the Geographic field.

superstition); d) the expectation with respect to the listener (closing the eyes while mentally making a wish). Formulaic expressions include pragmatemes (Mel'čuk, 1995; Blanco, 2013), stereotypical speech acts (Kaufer, 2018) and productive structures (Ovejas, 2022).

- d. Semantic relations not only include synonymy, antonymy and hyponymy, but also some productive relations not utilized so far in other dictionaries such as: a) "of an animal", which allows us to reflect the relationship between *hand/claw*, *mouth/snout*, *lip/beak*, *arm/thigh*, etc.; b) "profession or work", that shows the relationship between pairs of words such as *hand/chiropractor*, *feet/runner*, *bread/baker*, *wood/carpenter*, etc.
- e. Collocations are collected by Lexical Functions. This means that, considering a simple word, such as *hand*, *Diretes* offers all the verbs that combine with *hand* organized in a semantic way: a) verbs expressing the idea that 'to cause (the hand) to not work anymore', such as *tie*, *handcuff*; b) verbs expressing the idea 'it does not work properly', such as *get stiff*, *shake*; c) verbs expressing actions with a hand in a figurative sense, such as *hold out your hand*, *offer your hand*, etc.
- f. Idioms are collected following both onomasiological and semasiological approaches: for instance, *malas lenguas* ('gossiping tongues, evil tongues') are attached to the noun *falsedad* ('falseness'); similarly *becerro de oro* ('golden calf') and *buenas palabras* (lit. good words). Consequently, if someone does not remember an idiom related to *falsedad*, this person can go from the *falsedad* entry to consulting all entries of the three idioms.

2. Definitions

When writing definitions in the dictionary we considered two preliminary ideas. The first is the habitual risk of circularity in definitions. In order to avoid this, we were inspired by a previous work which attempts to show the advantages of the so-called *minimal and inverse definitions*: "*minimal* because I used to write only the minimal set of features necessary to understand the meaning of the word, and *inverse* because I wanted people to be able to decipher the word when reading the meaning" (Barrios, 2019, p. 196). This kind of definition must be written within the Natural Semantic Metalanguage model (NSM) (Goddard, 2012; 2016; Goddard & Wierzbicka, 2014; Wierzbicka, 1985; 1996). In particular, we resorted to two of their most relevant tools: primes (universal and simple words that cannot be decomposed) and molecules (universal words that can be explained by primes). As the methodological requirement for both tools exceed the temporal possibilities of our project in *Diretes* we did not use primes and molecules but simple words, meaning words that allow us to avoid circularity (for instance, if we write that a *table* is an *object used to put things, to work or to eat*, we consider that the meaning of the words *object*, *to use*, *to put*, *thing*, *to work* and *to eat* do not include the meaning of the word *table*).

The second point regarding definitions relates to the fact that we consider the meaning of a word includes not only semantic features but also world knowledge features: "I

no longer believe, again contrary to an age-old doctrine in word semantics, that it is possible and useful to separate core meaning from residual meaning and to consider the core as more important” (Kövecses, 1993, p. 247). As cultural features and data collected from referents can be relevant when attempting to understand the meaning of many words, we made the decision to divide the *Diretes*’s definitions into two parts: the first being a linguistic definition (written with basic words, as explained above); the second attempting to reflect cognitive links caused by extra-linguistic factors. We aim to illustrate both parts of these definitions in § 2.1 and § 2.2.

2.1 Definitions of Simple Words, Compound Words, and Idioms

Simple words, compound words, and idioms are described by means of minimal and expanded definitions, as explained above.

Illustration 1 shows both parts of the definition for the idiom *hombre anuncio* (sandwich-board man). Minimal definition (*definición mínima*) says (we translate to English the Spanish definition): “Person announcing something with a sign on the chest and another one on the back.” In this case, the minimal definition is considerably longer because this is the only way we found to present the meaning of this expression: it is necessary to include the features of *person*, *sing*, *chest* and *back*. Expanded definition (*definición expandida*) says: “This job is more than two centuries old, although in some cities, such as Madrid, it is prohibited (in this case, since 2008).” Accordingly, the second part of the definition adds some cultural features related to the use of this word in Spain (as *Diretes* describes the uses of Spanish from Spain):

Hombre-anuncio

Acepción núm 1:

Definición mínima: Persona que anuncia algo, con un cartel sobre el pecho y otro sobre la espalda.

Definición expandida: Es un trabajo con más de dos siglos de antigüedad, aunque en algunas ciudades, como Madrid, está prohibido (en este caso, desde 2008).

Etiqueta semántica: Persona con algún vínculo social o profesional.

Fig. 1: Definition for *hombre anuncio* (‘sandwich-board man’) in *Diretes*

Under the expanded definition, the entry contains the semantic label (*etiqueta semántica*), a type of hyperonym which belongs to a hierarchy of some nine levels of semantic labels. That particular hierarchy structures all the entries of *Diretes*. In this unequivocal case, the semantic label is “Person with a social or professional link.” Although it has not yet been applied in *Diretes*, the entries as in that for *hombre anuncio* will be illustrated through one picture to illustrate a more explicit meaning.

Before closing this section, it is necessary to mention a few important issues related to the definitions: a) the definitions have not been partially extracted automatically, although, in future work, we will attempt to automate the inheritance of shared semantic features; b) no particular corpora have been used to extract the encyclopaedic definitions, except for the data from Bosque (2004; 2006) and the sources mentioned in footnote 7.

2.2 Description for Formulaic Expressions

Formulaic expressions of *Diretes* are collected in a separate part of the database as they do not necessarily contain any meaning except that of a communicative intention. In fact, formulaic expressions portray information about a pragmatic function: for instance, when someone says *¡a buenas horas, mangas verdes!* (lit. ¡at good times, green sleeves!), the listener should understand that this expression is attached to the pragmatic function of reproaching: the speaker thinks the listener arrives too late to do what was supposed to be done, resembling an English speaker who says, in a similar situation, *too little too late!*

Regarding the formulaic expressions, the entry of *Diretes* contains a minimal and an expanded definition (in a similar way to the one explained in the §2.1). Minimal definition for this expression is translated to English: “A person who is in one place tells another person who is arriving at that time at that place that they are too late to do what they had to do together” (see Figure 2). Expanded definition adds some cultural information: “During the period of the Catholic Monarchs, there was a guard corps (called “Holy Brotherhood”) who wore green sleeved shirts as part of their uniform and had a reputation for being late when given notice.”

As this kind of expression does not have meaning, it makes no sense to assign a semantic label but a pragmatic function (such as to reproach, *reprochar*), as shown in Figure 2 (see *función pragmática*). In order to understand this function, these kinds of entries also offer a natural context with an example of use (see *situación típica* (typical situation) in Figure 2), for instance: “There is a family dinner. Several people in the family have arranged to meet a few hours before the dinner to prepare everything, but one of them arrives when they have already finished, and someone says to him: “A buenas horas, mangas verdes”.

¡A buenas horas, mangas verdes!

Acepción núm 1:

Definición mínima: Una persona que está en un lugar, le dice eso a otra persona que está llegando en ese momento para que entienda que llega demasiado tarde para hacer lo que tenían que hacer juntos.

Definición expandida: La expresión tiene su origen en la época de los Reyes Católicos, quienes reformaron un cuerpo, llamado Santa Hermandad, que tenía como objetivo perseguir a los delincuentes. Llevaban unas camisas con mangas verdes y solían llegar cuando ya se habían dispersado los causantes del altercado.

Función pragmática: Reprochar

Situación típica: Hay una cena familiar. Varias personas de la familia han quedado unas horas antes de la cena para prepararlo todo, pero uno de ellos llega cuando ya han terminado de preparar toda la cena, y le dicen: ¡A buenas horas, mangas verdes!

Fig. 2: Entry for the expression ¡A buenas horas, mangas verdes! (equivalent to Too little too late)

3. Semantic Relations

Semantic Relations in *Diretes* collect more than the usual relations (hyponymy, synonymy, etc.) As has been revealed, LFs do not cover all the semantic relations, hence we created new functions for them. Figure 3 illustrates this concept with a small group of words associated to *hand*:

Relaciones Semánticas					
	Id-FA	Id-Argumento	Id-Valor		He
+	_ABROGATIO	mano (s. f. sg.) 1 - Extremidad superior	muñón (s. m. SA) 1 - Sin asignar		No
+	_ABROGATUS	mano (s. f. sg.) 1 - Extremidad superior	manco (adj. c. c.) 1 - Defectuoso		No
+	_AFIJOIDE	mano (s. f. sg.) 1 - Extremidad superior	quiro- (morf. -) 1 - Parte de una palabra		No
+	_ANIMAL	mano (s. f. sg.) 1 - Extremidad superior	garra (s. m. sg.) 1 - Parte de un animal		No
+	_ANIMAL	mano (s. f. sg.) 1 - Extremidad superior	pezuña (s. m. sg.) 1 - Parte de un animal		No
+	_ANIMAL	mano (s. f. sg.) 1 - Extremidad superior	zarpa (s. m. sg.) 1 - Parte de un animal		No
+	_ARS	mano (s. f. sg.) 1 - Extremidad superior	prestidigitación (s. f. SA) 1 - Ocupación		No
+	_ARS	mano (s. f. sg.) 1 - Extremidad superior	quiropática (SA f. sg.) 1 - Profesión		No
+	_ARTIFEX	mano (s. f. sg.) 1 - Extremidad superior	prestidigitador (adj. c. c.) 1 - Persona		No
+	_ARTIFEX	mano (s. f. sg.) 1 - Extremidad superior	quiropático (s. m. sg.) 1 - Profesional		No
+	_AUCTUS	mano (s. f. sg.) 1 - Extremidad superior	manaza (s. f. SA) 1 - Sin asignar		No
+	_COMPOSITUM	mano (s. f. sg.) 1 - Extremidad superior	maniar (v. -) 1 - Sin asignar		No
+	_COMPOSITUM	mano (s. f. sg.) 1 - Extremidad superior	maniblanco (adj. c. c.) 1 - Animal		No
+	_COMPOSITUM	mano (s. f. sg.) 1 - Extremidad superior	manicorto (adj. c. c.) 1 - Cualidad		No
+	_COMPOSITUM	mano (s. f. sg.) 1 - Extremidad superior	manigordo (s. m. sg.) 1 - Animal		No
+	_COMPOSITUM	mano (s. f. sg.) 1 - Extremidad superior	manilargo (adj. c. c.) 2 - Cualidad		No
+	_COMPOSITUM	mano (s. f. sg.) 1 - Extremidad superior	maniostrar (v. -) 1 - Acción humana		No
+	_COMPOSITUM	mano (s. f. sg.) 1 - Extremidad superior	manirroto (adj. c. c.) 1 - Cualidad		No
+	_COMPOSITUM	mano (s. f. sg.) 1 - Extremidad superior	manivacio (adj. c. c.) 1 - Cualidad		No
+	_CORPORIS CURA	mano (s. f. sg.) 1 - Extremidad superior	manicura (s. f. SA) 1 - Sin asignar		No

Fig. 3: Some of the semantic relations related to *mano* (hand)

As there is a productive relation, related to “(like the one) of an animal”, we created a function called ANIMAL,¹⁰ which allows us to show the relation between *hand/claw*, *mouth/snout*, *lip/beak*, *arm/thigh*, etc. There is also a productive semantic relationship attached to something that is cancelled, which was labelled as ABROGATIO, meaning ‘abolition’, and covers, amongst other things, the relationship between *mano/muñón* (‘hand/stump of the arm’); ABROGATUS (‘who was cancelled’), covers the relationship between *mano/manco* (‘hand/one-handed’). ARTIFEX (meaning ‘artist’, ‘craftsman’, ‘expert’) covers the relationship between *hand/chiropractor*, *prestidigitator*, and ARS (‘art’) is used for the relationship between *hand/chiropractic*, *prestidigitation*.

Although Figure 3 only shows a small part of the set of these functions, to date a total of only ninety-seven functions naming semantic relations have been generated. Lack of space prevents us from presenting a complete list, however we mention some other functions such as A₀POSSESSOR (adjective for the owner of something), which covers *left-handed*, *right-handed*, *ambidextrous* for *hand*; A₁VITIUM (adjective for the first argument of something defective) such as *albino*, *hairless* for *hair*; CORPORIS CURA (body care) such as *manicure* for *hand* and *pedicure* for *foot*; CAEDO (trying to kill) such as *hang*, *decapitate* and *strangle* for *neck*, etc.

4. Collocations

Collocations of the dictionary are formalized in the usual manner within the MTT, that being by means of LFs. As mentioned before, some of our collocations were obtained automatically (which is a novelty in the MTT scenario, for further information see Barrios, 2010): we will try to illustrate it in the following lines.

As we work with lexical fields, we can predict some of the regular collocations shared by co-hyponyms, such as the words for joints: *elbow*, *wrist*, *knee* and *ankle*. As these words share semantic features, we were able to incorporate some collocations attached to the following LFs:

- a) FinReal₁, which means ‘not to do work anymore’ and covers the verbs *dislocar/desencajar* (‘to dislocate’), *fracturarse* (‘to fracture’), *romperse* (‘to break’);
- b) IntentarCaus₁PlusBonFact₀, which means ‘(the first argument of the verb) try in order that something works in a better way’, and covers the verb *operarse* (‘to operate on’);
- c) IntentarCausPlusBonFact₀, which means ‘to try to make something work in a better way’ and covers the verb *escayolar* (‘to put a cast on’) and *inmovilizar* (‘to immobilize’);
- d) S₀AntiBonFact₀, which means ‘noun that expresses that something does not work properly’ and covers the nouns *esguince* (‘sprain’), *lesión* (‘lesion’), *torcedura* (‘strain’), *fractura* (‘fracture’), etc.

¹⁰ In order to distinguish the functions between the LFs of the MTT, we use a capital letter: our non-canonical LFs are functions we created *ad hoc* to facilitate the retrieval of information in an onomasiological sense. We also chose Latin words, as MTT researchers made with LFs.

Figure 4 shows some of these collocations related to *tobillo* ('ankle') (the last column portrays which collocations were obtained automatically through a yes/no marc):

Relaciones Semánticas				
Id-FA	Id-Argumento	Id-Valor	Here	
⊕ _MORBUS	tobillo (s. m. sg.) 1 - Articulación	sinovitis (s. f. SA) 1 - Sin asignar	Sí	
⊕ _PROTECTOR	tobillo (s. m. sg.) 1 - Articulación	tobilera (s. f. SA) 1 - Sin asignar	No	
⊕ AOAntiBonFact0	tobillo (s. m. sg.) 1 - Articulación	hinchado (adj. c. c.) 1 - Sin asignar	Sí	
⊕ AOAntiBonFact0	tobillo (s. m. sg.) 1 - Articulación	inflamado (adj. c. c.) 1 - Sin asignar	Sí	
⊕ AOAntiBonFact0	tobillo (s. m. sg.) 1 - Articulación	lesionado (adj. c. c.) 1 - Sin asignar	Sí	
⊕ AntiBonFact0	tobillo (s. m. sg.) 1 - Articulación	hinchar (v. - -) 1 - Sin asignar	Sí	
⊕ AntiBonFact0	tobillo (s. m. sg.) 1 - Articulación	inflamar (v. - -) 1 - Sin asignar	Sí	
⊕ AntiBonFact1	tobillo (s. m. sg.) 1 - Articulación	doler (v. - -) 1 - Sin asignar	Sí	
⊕ AntiBonReal1	tobillo (s. m. sg.) 1 - Articulación	entumecer (v. - -) 1 - Sin asignar	Sí	
⊕ AntiBonReal1	tobillo (s. m. sg.) 1 - Articulación	lesionar (v. - -) 1 - Sin asignar	Sí	
⊕ AntiBonReal1	tobillo (s. m. sg.) 1 - Articulación	torcer (v. - -) 1 - Acción	No	
⊕ CausPredPlusBon	tobillo (s. m. sg.) 1 - Articulación	fortalecer (v. - -) 1 - Sin asignar	Sí	
⊕ FinReal1	tobillo (s. m. sg.) 1 - Articulación	desencajar (v. - -) 1 - Sin asignar	Sí	
⊕ FinReal1	tobillo (s. m. sg.) 1 - Articulación	fracturar (v. - -) 1 - Sin asignar	Sí	
⊕ FinReal1	tobillo (s. m. sg.) 1 - Articulación	romper (v. - -) 1 - Sin asignar	Sí	
⊕ IntentarCaus1PlusBonFact0	tobillo (s. m. sg.) 1 - Articulación	operar (v. - -) 1 - Sin asignar	Sí	
⊕ IntentarCausPlusBonFact0	tobillo (s. m. sg.) 1 - Articulación	escayolar (v. - -) 1 - Sin asignar	No	
⊕ IntentarCausPlusBonFact0	tobillo (s. m. sg.) 1 - Articulación	inmovilizar (v. - -) 1 - Sin asignar	Sí	
⊕ Parte de	tobillo (s. m. sg.) 1 - Articulación	articulación (s. f. sg.) 1 - Estructura	No	
⊕ Real1	tobillo (s. m. sg.) 1 - Articulación	doblar (v. - -) 1 - Sin asignar	Sí	
⊕ SOAntiBonFact0	tobillo (s. m. sg.) 1 - Articulación	dolor (s. m. sg.) 1 - Sentimiento negativo	No	
⊕ SOAntiBonFact0	tobillo (s. m. sg.) 1 - Articulación	esguince (s. m. SA) 1 - Sin asignar	Sí	
⊕ SOAntiBonFact0	tobillo (s. m. sg.) 1 - Articulación	lesión (s. f. sg.) 1 - Efecto físico en animados	No	
⊕ SOAntiBonFact0	tobillo (s. m. sg.) 1 - Articulación	torcedura (s. f. SA) 1 - Sin asignar	No	
⊕ SOFinReal1	tobillo (s. m. sg.) 1 - Articulación	fractura (s. SA SA) 1 - Sin asignar	Sí	
			No	

Fig. 4: Some of the collocations related to *tobillo* (ankle)

5. Idioms

Idioms have their own entry in *Diretes*. The novelty being, thanks to the LFs and the functions mentioned in the § 4, idioms are accessible both by semasiological and onomasiological senses. Hereinafter, we illustrate this by an example of a concrete case, the Spanish idioms related to *false* and *falseness*.

The first one is an adjective associated with certain idioms, such as *figura decorativa* ('figurehead'), *caballo de Troya* ('Trojan horse'), *castillo de naipes* ('house of cards'), *cuento chino* (lit. Chinese tale, meaning tall story), *flor de un día* (lit. flower of one day, meaning short-lived), *música celestial* (lit. heavenly music, meaning music to my ears), *un quiero y no puedo* (lit. a want and I can't, meaning will o' the wisp). We applied several grammatical tests to verify that all these idioms function as an adjective (see Barrios, 2022a) and then we associated these idioms with the adjective *falso* ('false'). In this way, if someone does not remember some of these idioms, they can consult the entry *falso* and all the entries specific to each idiom will be accessible by clicking.

The semasiological sense is present, as always, when writing the idiom in the search engine of the dictionary. Instead of *false*, the user can write *falsedad* ('falseness'). The idioms related to this word function as a noun, therefore *Diretes* will offer information about *becerro de oro* ('golden calf'), *buenas palabras* (lit. good words, meaning nice but false words) and *malas lenguas* (lit. bad tongues, meaning gossip, added a nuance of falsehood).

Figure 5 shows the results for *falso* ('false'):

falso				
Palabras relacionadas				
Aceptación núm: 1				
Definición mínima: Que no es lo que se podría pensar que es.				
Definición expandida: Se dice de cosas (piel falsa), de palabras (lo que dices es falso), de resultados (el resultado de la suma es falso), etc.				
Etiqueta semántica: Cualidad				
falso como argumento				
Locuciones nominales				
Propiedad de personas en relación con personas	falso	figura decorativa	nivel ELE: S	
Propiedad de personas, entidades o situaciones	falso	caballo de Troya	nivel ELE: S	
Propiedad de personas, entidades o situaciones	falso	castillo de naipes	nivel ELE: S	
Propiedad de personas, entidades o situaciones	falso	cuento chino	nivel ELE: S	
Propiedad de personas, entidades o situaciones	falso	flor de un día	nivel ELE: S	
Propiedad de personas, entidades o situaciones	falso	música celestial	nivel ELE: S	
Propiedad de personas, entidades o situaciones	falso	un quiero y no puedo	nivel ELE: S	

Fig. 5: Idioms related to *falso* (false)

6. Conclusions

Diretes is an ongoing project with considerable information still to be added as well as problems to be solved. The majority of the difficulties found relate to the following aspects: a) correctly writing a minimal definition with simple words; b) selecting the most appropriate pragmatic function for formulaic expressions; c) avoiding overproduction of functions created *ad hoc* for semantic relations; d) selecting the correct LFs for collocations; e) selecting the most accurate words (often more than one) for the onomasiological search of idioms.

Other problems are related to questions that were impossible for us to cover in this paper, such as the selection of the examples appropriate to different levels of Spanish learning, or the debugging of errors, a very important task in *Diretes* due to the automatization of the procurement of collocations and the LFs attached to them.

Although many years away from being completed, we aspire in the long run to be able to refine the working methodology in order to improve its efficiency.

References

- Alonso Ramos, M. (2004). Elaboración del Diccionario de Colocaciones del español y sus aplicaciones. In P. Bataner, & J. de Cesaris (Eds.), *De lexicografía. Actes del I Simposium Internacional de Lexicografía* (pp. 149–162). IULA & Edicions Petició.
- Apresjan, J. (1974). Regular polysemy. *Linguistics* 142/5, 5–32.
- Apresjan, J. (1992). Systemic Lexicography as a Basis of Dictionary-making. *Dictionaries: Journal of the Dictionary Society of North America* 14, 79–87.
- Apresjan, J. (2000). *Systemic Lexicography*. Oxford University Press.
- Apresjan, J. (2009). The Theory of Lexical Functions: An Update. In D. Beck et al. (Eds.), *Fourth International Conference on Meaning-Text Theory* (pp. 1–14). Observatoire de linguistique Sens-Texte.
- Apresjan, J. (2018). *New Explanatory Dictionary of Russian Synonyms*. PubMix.com.
- Apresjan, V. (2011). Active dictionary of the Russian language: theory and practice. In I. Boguslavsky, & L. Wanner (Eds.), *Proceedings of the 5th International Conference on Meaning-Text Theory* (pp. 13–24). Universitat Pompeu Fabra.
- Atkins, S. B. T., & Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford University Press.
- Barrios, M. A. (2010). El dominio de las Funciones Léxicas en el marco de la Teoría Sentido-Texto. *Estudios de Lingüística del Español* 30, 1–477.
- Barrios, M. A. (2010). Minimal and inverse definitions: a semi-experimental proposal when building a Spanish dictionary with semantic primes and molecules. In P. Bert, K. Mullan, & L. Sadow (Eds.), *Studies in Ethnopragmatics, (Cultural) Semantics, and Minimal English* (pp. 191–212). Springer.
- Barrios, M. A. (2020). The making of the Diretes dictionary. In Z. Gavrilidou, M. Mitsiaki, & A. Fliatouras (Eds.), *Lexicography for Inclusion. EURALEX Proceedings. XIX Congress of the European Association for Lexicography* (pp. 13–22).
- Barrios, M. A. (2022a). *DiRetEs*: Diccionario Reticular del Español. Resultados de la primera fase de un proyecto. In L. Iomdin, J. Milićević, & A. Polguère (Eds.), *Lifetime Linguistic Inspirations. To Igor Mel'čuk from Colleagues and Friends for his 90th Birthday. Wiener Slawistischer Almanach Sonderband* 101 (pp. 73–84).
- Barrios, M. A. (2022b). Meaning and function of Spanish formulemes and pragmatemes vs. illocutionary verbs. *Russian Journal of Linguistics* 26/4, 1031–1049. doi: 10.22363/2687-0088-31597
- Blanco, X. (2013). Les pragmatèmes: définition, typologie et traitement lexicographique. *Verbum* 4, 17–25.
- Frawley, W. (1997). New forms of specialized dictionaries. *International Journal of Lexicography* 1(3), 189–213.

- Goddard, C. (2012). Semantic Primes, Semantic Molecules, Semantic Templates: Key Concepts in the NSM Approach to Lexical Typology. *Linguistics* 50(3), 711–743. <https://doi.org/10.1515/ling-2012-0022>
- Goddard, C. (2016). Semantic molecules and their role in NSM lexical definitions. *Cahiers de lexicologie*, 109, 13–34.
- Goddard, C., & A. Wierzbicka (2014). *Words and Meanings. Lexical Semantics across Domains, Languages and Cultures*. Oxford University Press.
- L’Homme, M. C. (2012). Using ECL (Explanatory Combinatorial Lexicology) to discover the lexical structure of specialized subject fields, In Apresjan, J. et al. (Eds.), *Words, Meanings and other Interesting Things. A Festschrift in Honour of the 80th Anniversary of Professor Igor Alexandrovic Mel’cuk* (pp. 378–390). RCK.
- L’Homme, M. C. (2018). Maintaining the balance between knowledge and the lexicon in terminology: a methodology based on frame semantics. *Lexicography* 4(1), 3–21.
- L’Homme, M. C., B. Robichaud, & C. Subirats. (2016). A Proposal for combining “general” and specialized frames. In *Proceedings of the Workshop on Cognitive Aspects of the Lexicon* (pp. 156–165).
- Kauffer, M. (2018). Qu’est-ce qu’un ALS ? *Verbum* XL/1, 35–50.
- Mel’cuk, I. (1984). Le Dictionnaire Explicatif et Combinatoire du Français Contemporain. Prol. *Dictionnaire Explicatif et Combinatoire du Français Contemporain. Recherches lexicosémantiques I* (pp. 3–15). Les Presses de l’Université de Montréal.
- Mel’cuk, I. (1995). Phrasemes in Language and Phraseology in Linguistics. In E. Everaert, J. van der Linden, A. Schenk, & R. Schreuder (Eds.), *Idioms. Structural and Psychological Perspectives* (pp. 167–232). Lawrence Erlbaum Associates.
- Mel’cuk, I. (1996). Lexical Functions: A Tool for the Description of Lexical Relations in Lexicon. In L. Wanner (Ed.), *Lexical Functions in Lexicography and Natural Language Processing* (pp. 37–102). John Benjamins.
- Mel’cuk I. (2014). *Semantics. From meaning to text*. Volume 3. John Benjamins.
- Mel’cuk I. (2023). *General Phraseology. Theory and Practice*. John Benjamins.
- Mel’cuk, I., Clas, A., & Polguère, A. (1995). *Introduction à la lexicologie explicative et combinatoire*. Editions Duculot.
- Mel’cuk, I., & Polguère, A. (2007). *Lexique actif du français – l’apprentissage du vocabulaire fondé sur 20.000 dérivations sémantiques et collocations du français*. Boeck & Larcier.
- Mel’cuk, I., & Polguère, A. (2018). Theory and Practice of Lexicographic Definition. *Journal of Cognitive Science* 19/4, 417–470. 10.17791/jcs.2018.19.4.417
- Polguère, A. (2000). Towards a theoretically-motivated general public dictionary of semantic derivations and collocations for French. In U. Heid, S. Evert, E. Lehmann, & C. Rohrer (Eds.), *Proceedings of the Ninth EURALEX International Congress* (pp. 517–527). Stuttgart.

Polguère, A. (2003). Collocations et fonctions lexicales: pour un modèle d'apprentissage. In F. Grossmann, & A. Tutin (Eds.), *Les Collocations: analyse et traitement* (pp. 117–141). Éditions de Werelt.

Polguère, A. (2007). Lexical Function Standardness. In L. Wanner (Ed.), *Selected Lexical and Grammatical Issues in the Meaning-Text Theory. In Honour of Igor Melčuk* (pp. 43–96). John Benjamins.

Polguère, A. (2014). From writing dictionaries to weaving lexical networks. *International Journal of Lexicography* 27/4, 396–418. <https://doi.org/10.1093/ijl/ecu017>

Roche, C. (2015). Ontological definitions. In H. J. Kockaert, & F. Steurs (Eds.) *Handbook of Terminology: Volume 1* (pp. 128–152). John Benjamins.

Wierzbicka, A. (1985). *Lexicography and Conceptual Analysis*. Karoma.

Wierzbicka, A. (1996). *Semantics: Primes and Universals*. Oxford University Press.

Acknowledgements

This work was carried out within the framework of the project DiRetEs (Second Phase) and was supported by the Ministry of Science and Innovation of Spain (PID2021-122894-NB-I00) and the FEDER Program (European Union). I would like to express my gratitude to the anonymous reviewers for their helpful observations, to Svetlana Antropova and Deborah Paton for all her suggestions on the English revision of the manuscript and to Begoña Sanromán for her outstanding help. Any remaining errors are the sole responsibility of the author.

Contact information

María Auxiliadora Barrios Rodríguez

Universidad Complutense de Madrid

auxibarríos@filol.ucm.es

Luke Omoyemi Akinremi and María José Domínguez Vázquez

THE ERASMUS MUNDUS JOINT MASTER IN LEXICOGRAPHY – EMJM-EMLEX New Developments and Goals

Abstract This paper introduces the Erasmus Mundus Joint Master in Lexicography – EMJM-EMLex - especially some new developments and objectives. The EMJM-EMLex programme remains focused on lexicography, but is evolving into a multidisciplinary, digital discipline in order to adapt to societal and scientific changes. Founded in 2009, experience to date confirms that the success of EMLex lies in its unique objectives and in responding to the ever-changing needs of society while fostering international collaboration, linguistic diversity and cultural exchange (Dias, 2022). The programme’s multidisciplinary approach is in line with the Council of Europe’s language policy goals, emphasising lifelong learning and multilingualism. With 8 full member universities and 8 associate members in its consortium, the programme offers an international and interdisciplinary curriculum covering lexicography, digital humanities, language technologies and natural language processing. EMJM-EMLex integrates theoretical and practical aspects of lexicography over four semesters, including mobility periods, advanced modules, internships and a master’s thesis, after which graduates receive a joint degree from the member universities. EMJM-EMLex graduates are prepared for diverse career opportunities. The programme’s commitment to innovation, internationalisation and sustainability ensures its continued relevance and impact in the digital age.

Keywords lexicography; digital humanities; language technologies; EMJM-EMLex

1. Introduction

Given the challenge of bridging the gap between theoretical and practical lexicography, there had been contemplations of the possibility of establishing a lexicography master’s degree programme since the 1980s; which however were typically met with initial reluctance and a pessimistic outlook, partly due to the perceived low demand for lexicographers, while overlooking the evolving needs of modern lexicographic products and the growing importance of online lexicography (Schierholz, 2010, pp. 343-344). Overcoming various difficulties, the European Master in Lexicography (EMLex) has been offered since 2009. This international master’s degree (120 ECTS) has been awarded the title of “Erasmus Mundus Joint Master Degree” by the Education, Audiovisual and Culture Executive Agency since 2016.

Through its broad curriculum, EMLex has been training the next generation of lexicographers who are versed and well established in both lexicographic theory and practice, and are suited to tackle modern language and lexicographic problems. Therefore, the programme offers a globally oriented and interdisciplinary master’s

degree programme in lexicography aimed at students who have an interest in engaging with dictionaries and digital lexicographic tools.

For a third time, the Audiovisual and Culture Executive Agency has granted the recognition of Erasmus Mundus Programme to EMLex for the period of 2024–2030. An increasing interdisciplinarity of the programme, EMJM-EMLex, was proposed for this new period, incorporating aspects such as digital humanities, statistics, artificial intelligence, language technologies, and natural language processing. Thus, students will be equipped with the relevant competencies to meet the linguistic and lexicographic demands of an evolving digital and globalised society.

The new EMJM-EMLex consortium consists of eight full member universities and eight associate members. The full members include Universidade de Santiago de Compostela (coordinating institution), Università degli Studi Roma Tre, Universidade do Minho, Friedrich-Alexander-Universität Erlangen-Nürnberg, Károli Gáspár Református Egyetem, Université de Lorraine, Ilia State University, and Universität Hildesheim. The associated members are Stellenbosch University, Leibniz-Institut für Deutsche Sprache, University of Surrey, Akademie der Wissenschaften zu Göttingen, Université Félix Houphouët-Boigny, Universität Innsbruck, Pontificia Universidad Católica de Valparaíso, and the University of Texas at Austin.

2. New Developments and Goals

2.1 New Countries in Focus

One of the new developments in EMJM-EMLex is the focus on Africa and South America. In this international master programme, efforts are being made to reach a wider global audience, and this includes promoting the programme beyond Europe. For the first intake, a focused social media advertising campaign is being launched to bolster the brand presence of EMJM-EMLex, to create more awareness about the programme and to attract young and motivated future lexicographers especially from Africa¹. In those parts of the world, there are many languages that have not received much attention in terms of lexicographic development and the creation of lexicographic resources for speakers of such languages. Thus, there is huge potential to be harnessed by training lexicographers from those continents.

In addition, this campaign enabled the observation of the interest that lexicography awakens, and indeed, it aroused a lot of interest. The advertising campaign was launched in early 2024, during the application period for the 2024/25 academic year. Targeting individuals aged 20 to 35 with bachelor's degrees in language-related fields across various African cities in Nigeria, Cameroon, Togo, Egypt, South Africa, Kenya, Namibia, Senegal, Ivory Coast, Morocco, Ghana, Tanzania, and Ethiopia, the campaign utilised Facebook and Instagram as primary platforms. It featured a 15-second promotional video released from February 5th to 29th, followed by newsfeed and stories ads from February 12th to 29th.

¹ For the next intake, South America will be the target continent.

The campaign garnered significant traction, amassing 3,779,263 impressions with a commendable 57% View Through Rate (VTR), indicative of strong engagement among the target audience. Moreover, the Click Through Rate (CTR) of 0.72% surpassed expectations, signifying effective targeting of users interested in pursuing the master's programme. Notably, the campaign yielded its most favourable outcomes in Ghana, South Africa, and Ivory Coast, underscoring the efficacy of the advertising efforts in the continent. This interest was also reflected in the number of applicants.

2.2 Lexicography as a Multidisciplinary Discipline

The EMJM-EMLex programme is undergoing transformative developments to adapt to the evolving landscape of digitalisation and societal needs. While the core focus remains on lexicography, the programme is evolving into a multidisciplinary digital discipline, aligning with the demands of a global and multilingual society. In response to the increasing emphasis on multilingualism and lifelong learning by organisations such as the Council of Europe, there is a growing need for critical approaches to lexical information systems (see Villa Vigoni Theses on Lexicography of EMLex, 2018). Moreover, the market for lexicographic products is very large and, as Granger has already pointed out “barriers between the different types of language resources – dictionaries, encyclopedias, term banks, lexical databases, vocabulary learning tools, writing aids, translation tools – are breaking down.” (Granger, 2012, p. 4). Furthermore, we agree with Gouws (2017, p. 26) that in today's world neither lexicography nor lexicographical didactics can deal only with dictionaries – even if they are electronic – but must focus on the whole range of resources and media used, especially by the younger generation for these purposes. Taking this into account, new technologies and methods of analysis, as well as studies on users *in actu* (Wiegand, 1988, p. 501), can contribute to the design of better resources.

The social changes are also accompanied by the increasing intertwining of lexicography with natural language processing (NLP) and artificial intelligence (AI). Lexicographical knowledge is central to the development of resources for humans or machines at a wide variety of levels. It supports the responsibility not only of conveying linguistic, specialised and cultural knowledge, but also of the specialised conversion and application of this knowledge in modern user-oriented reference works and resources. In this context, we understand lexicography as a link between lexicographical expertise, AI and the digital humanities.

Internally, the programme is tasked with addressing subject-related tasks such as establishing high-quality knowledge repositories, such as reference works, dictionaries, databases, information systems, as well as training conditions to ensure the reliable data supply in modern information societies. Externally, in the era of big data, machine translation systems, robots for automatic text production, digital humanities, as well as regarding new methods and techniques for automatic data collection and language generation, lexicography fulfils the task of a mediating authority between different worlds of knowledge and research, and between the world of research, the academic world and society.

The aforementioned requires the integration of current scientific research results and innovations into the curriculum, which also contributes to enhancing students' key digital competences and skills, for example:

- NLP-related content is offered, e.g., in courses such as *Basics of Lexicography* (1st semester) and in advanced seminars such as *Computational Lexicography* or *Modelling and Representing Data in Digital Lexicography* (2nd semester).
- Machine Translation, automatic translation and generation, as well as the role of AI for gaining lexicographical data are topics in courses such as *Dictionaries and Translation* (2nd Semester) and in the in-depth modules in the 3rd semester.
- Statistical content and visualisation of digital data as well as languages and technologies are taught, e.g., in the 3rd semester and in special courses.
- The interaction of new analysis and data retrieval strategies, the emergence of new lexicographic resources and the evolution of dictionary use requires knowledge to establish quality criteria and resources appropriate to users. Advanced seminars such as *Metalexicography* and *Research into Dictionary Use* play a key role here. Metalexicography and dictionary use research are areas that are not only important on a theoretical level; the last-mentioned field, for example, is becoming increasingly central to adequate media literacy, as required in many (school) educational guidelines in European countries.

EMJM-EMLex distinguishes itself by offering a multilingual context for acquiring multidisciplinary skills, particularly focusing on lexicographic and lexical knowledge. Through interaction with programme stakeholders, students are exposed to a range of courses that cater to their diverse cultural and social backgrounds while meeting international educational objectives for lexicography. This unique approach ensures that EMJM-EMLex graduates are equipped to analyse data effectively, develop innovative technologies, and critically engage with evolving modes of communication in today's globalised world.

2.3 Lexicography and Society

The first dictionaries ever produced may seem primitive according to the present standard, but their authors at least had the privilege of spontaneously understanding the social value of their work, i.e., the close relation between specific types of social needs and the solutions given by means of dictionaries. With the passing of the centuries and millennia, this close relation was forgotten. [...] The social needs originally giving rise to lexicography were relegated to a secondary plane and frequently ignored (Tarp, 2009, p. 19).

Perhaps today, more than ever, it is necessary to make society aware of the value of lexicography. In a global society of immediacy and easy access to a large amount of data with a click, how can a user know that the resource they are handling is reliable?

The master's programme addresses these kinds of questions, but goes a step further by trying to involve society in lexicographic research and discussion. For example:

- Citizens can participate in some collaborative lexicographic projects; this is a lexicography for the citizens with the citizens, an open science lexicography.
- The programme fosters societal involvement through events such as EMLex Hackathons and Congresses, as well as proposes dissemination actions through TV-media, social networks focusing on language, lexicography, media literacy, etc.
- Outreach to educational institutions and the offer of training seminars for teachers and lecturers with a special focus on new digital resources is important in two senses: on the one hand, due to the necessary updating from a content or technical point of view for the purpose of using new resources in teaching and critical media competence; and on the other hand, because of the necessary sensitisation regarding the need for high quality lexicography, which in turn is fundamental for success in other professions, e.g., translators, interpreters, etc.
- The involvement of society also implies an approach based on the identified professional, social, institutional and industrial (market-related) needs. It is the task of the Consortium to identify gaps in the job market and create job networks. But it is also the task of the students and graduates to establish networks, not only with the entire teaching staff of the Master's programme, but also during their internships in centres all over the world.

In addition, it is very important to emphasise the internationalisation of the programme in order to ensure the exchange of experiences for the further development of institutional and personal cooperation networks with special attention to a multilingual, multicultural and egalitarian society and to maintain and intensify the established structures in the international professional world for students and teachers.

3. Conclusion

The specific development of optimisation actions within the EMJM-EMLex programme focuses on meeting professional needs and increasing the visibility of lexicography as a multidisciplinary field in the digital humanities. EMJM-EMLex faces the challenge of rapidly evolving demands in lexicography and dictionary research and strives to adapt itself constantly to meet these demands. This is exemplified by the recent expansion of the programme to include disciplines such as digital humanities, statistics, artificial intelligence and natural language processing, ensuring that graduates are well equipped to tackle the complexities of modern lexicography.

By extending its reach to Africa and South America, EMJM-EMLex not only broadens its geographic scope but also addresses the underrepresentation of certain languages in lexicographic resources. The success of targeted social media campaigns highlights

the significant interest in lexicography in these regions, suggesting a promising future for the inclusion of diverse linguistic perspectives in global lexicographic work.

Furthermore, the societal impact of lexicography, a focus of EMJM-EMLex, is also a critical aspect. By involving citizens in collaborative projects and promoting outreach through events and training seminars, EMJM-EMLex not only enhances the visibility and understanding of lexicography but also fosters a community that values and supports the development of lexicographic resources.

The future of lexicography lies in offering metalexicographic standards both in print and online, analysing large volumes of data using new NLP and AI techniques, developing information systems according to user types and quality criteria, documenting language and cultural knowledge, and being a socially responsible, inclusive and citizen-friendly lexicography, and EMJM-EMLex aims to provide a framework that prepares future lexicographers to adequately meet these needs.

References

- Dias, I. (2022). European Master in Lexicography: A Platform for Scientific and Cultural Exchange and Dialogue: To the Carolina Michaëlis Class 2022, University of Minho, Braga, Portugal. *Lexicographica*, 38(1), 503–511. <https://doi.org/10.1515/lex-2022-0017>
- EMLex (2018). *Villa Vigoni Theses on Lexicography*. Retrieved May 5, 2024, from <https://www.emlex.eu.usc.gal/about-us/publications/other/>
- Gouws, R. H. (2017). La sociedad digital y los diccionarios. In M. J. Domínguez Vázquez, & M. T. Sanmarco Bande (Eds.), *Lexicografía didáctica* (pp. 17–34). Peter Lang.
- Granger, S. (2012). Introduction: Electronic lexicography – from challenge to opportunity. In S. Granger, & M. Paquot (Eds.), *Electronic Lexicography* (pp. 1–11). Oxford University Press.
- Schierholz, S. (2010). EMLex: Europäischer Master für Lexikographie – European Master in Lexicography. *Lexicographica*, 26(1), 343–350. <https://doi.org/10.1515/9783110223231.6.343>
- Tarp, S. (2009). Reflections on Lexicographical User Research. *Lexikos*, 19, 275–296.
- Wiegand, H. E. (1998). *Wörterbuchforschung: Untersuchungen zur Wörterbuchbenutzung, zur Theorie, Geschichte, Kritik und Automatisierung der Lexikographie*. De Gruyter.

Contact information

Luke Omoyemi Akinremi

Universidade de Santiago de Compostela
luke.akinremi@usc.es

María José Domínguez Vázquez

Universidade de Santiago de Compostela
majo.dominguez@usc.es

Dwayne Ellul

MALTESE LEXICOGRAPHY

A Historical Context and the Current State

Abstract This study has two main objectives: firstly, to provide a concise overview of the evolution of Maltese lexicography, and secondly, to shed light on the latest advancements in the field within the local context. Following select criteria for dictionary reviews presented in Svensén's (2009, pp. 482–483) checklist, this paper reviews the most significant dictionaries in the Maltese lexicographic tradition, starting from Thezan's dictionary, compiled in the early 17th century, and continuing with the most important ones from each subsequent century.

Keywords Maltese lexicography; trilingual lexicography; bilingual lexicography; monolingual lexicography; dictionary criticism; online dictionaries; root-and-pattern morphology; concatenative morphology

1. Introduction

Maltese is an ancient language that has a rich history spanning over a millennium. In a simplified manner, the stratigraphy of the language can be presented in terms of three different strata: the Semitic (Arabic) stratum, the Romance superstratum and the English adstratum (Mifsud, 1995, pp. 26–27). Maltese became recognised as an official language of Malta, alongside English, in 1934. Then, in 1964, when Malta gained independence and ceased to be a British colony, Maltese was also recognised as the national language. The Għaqda tal-Kittieba tal-Malti [Association of Writers of Maltese] standardised the alphabet and spelling in 1921 and 1924, respectively and were legally accepted by the State in 1934. However, Malta has a long-standing lexicographic tradition dating back hundreds of years.¹ In spite of this, Maltese rarely features in literature on lexicography. Given that it is a language that exhibits various linguistic systems in contact, it would be useful to provide an overview of its most important lexicographic resources.

The paper is structured as follows. Section 2.1 focuses on bilingual and trilingual dictionaries in which, as a result of the Maltese historical and cultural context, English does not feature as one of the languages. On the other hand, section 2.2 deals with dictionaries in which English does feature whereas section 2.3 deals with monolingual dictionaries. Section 3 provides insights about recent local advancements in the field, specifically two ongoing initiatives: Camilleri's *Ġabra*, an open online lexical database for Maltese, and *Ġabra Terminologika*: an online terminology resource currently being developed at the University of Malta.

¹ For a comprehensive list of Maltese dictionaries see <https://kunsillmalti.gov.mt/rizorsi/listabiblijografika/>

2. The Maltese Lexicographic Tradition

The first 9 Maltese words in a dictionary can be found in *Thesaurus Polyglottus* which Hieronymus Megiser published in Frankfurt in 1603 after visiting Malta in 1588 (Cassola, 1986, p. 21). One can observe the proliferation of Maltese words in 17th and 18th centuries vocabulary lists (Brincat, 2011, pp. 243–248). This study will not address such lists, as it will concentrate on the most significant dictionaries in the Maltese lexicographic tradition.

Following the norm of the early stages of lexicography in most languages, Maltese bilingual and multilingual dictionaries are far more common than monolingual ones. This is a result of cultural and historical factors. During the period between 1530 and 1798, when Malta was ruled by the Order of St John, the languages that featured in dictionaries alongside Maltese, were Latin and Italian: the Order's official languages. While Italian continued to be used in some dictionaries during the 19th century,² it was progressively replaced by English. The earliest monolingual dictionary and the two most comprehensive Maltese dictionaries—one bilingual and one monolingual—appear in the 20th century.

2.1 Bilingual and Trilingual Dictionaries Before English

The oldest dictionary in our possession is the *Regole per La Lingua Maltese* [*Rules for the Maltese Language*], a manuscript discovered by Cassola at the Biblioteca Vallicelliana in Rome. Cassola (1992, p. xxxiii) claims that the manuscript in question is a copy which Monsieur le Chevallier De La Bastie Fogasses based on a dictionary authored by the French Knight Thezan during the early 17th century. The Vallicelliana manuscript (hereinafter *Regole*), cannot be dated past 1719 (p. xxx).

Regole starts with a short seventeen-page grammar (Cassola, 1992, pp. 3r–11r) written in Italian. Next, there is the bidirectional bilingual Italian-Maltese (pp. 14r–112r) and Maltese-Italian (pp. 115r–185v) dictionary. With a few exceptions, the macrostructure of the first part is organised alphabetically using the Latin alphabet, whereas in the second part a combination of Latin and Arabic characters is used. As noted by Cassola (1992, p. xi), until approximately 1844 it was customary for a small number of individuals who wrote in Maltese to include Arabic graphemes for phonemes that were not represented in the Latin alphabet.

The Italian-Maltese section comprises 3,925 Italian lemma signs paired with their Maltese equivalents, with a further 5 (p. xlvii) lemmas to be included, bringing the total to 3,930. The Maltese-Italian section consists of 3,110 lemma signs, with an additional 4 lemmas (xlvi) that need to be included, bringing the total to 3,114.

² A significant dictionary is the *Vocabolario della Lingua Maltese* [*Vocabulary of the Maltese Language*], published by A.E. Caruana in 1903. This dictionary was the first to organise lemmata according to their root into nests, following the Arabic way of listing. It was also the first to provide etymological information, even though according to Saydon (24), its reliability is questionable. One significant critique of this dictionary is that Caruana excluded numerous contemporary non-Semitic words in Maltese in order to maintain the macrostructure's functionality. For critical comments about this dictionary see Saydon (pp. 23–24), Privitera (pp. 28–29), and Agius (pp. 174–175).

Alfieri (in Brincat, 2011, p. 250) states that the Italian-Maltese section does not rely on established Italian dictionaries of that era, and its synonyms seem to have been directly derived from practical usage. Furthermore, it seems that the lemma signs for both sections were taken directly from usage. This demonstrates that the primary objective of the dictionary was practical: to aid the Knights in effectively communicating with the monolingual Maltese population. Indeed, the manuscript concludes with a section titled “Ta ϵ lim a ϵ l Soldat” [Teaching a Soldier] which is a series of drill commands for training with weapons in Italian translated into Maltese.

A significant dictionary in the Maltese lexicographic tradition is the ***Ġamma tal Kliem Kartaginis mscerred fel Fom tal Maltin u Ghaucin*** [A collection of Carthaginian words spoken by the Maltese and the Gozitans], which was being compiled by Canon Gianfrancesku Agius Sultana Ghauci, also known as De Soldanis, before 1750 and completed around 1767. De Soldanis had intended to publish the dictionary (Agius De Soldanis, 1750, p. 19), but he passed away in 1770. As a result, the dictionary remained in manuscript form until 2016, when the first volume was finally published by Carabott. Prior to the discovery of the *Regole*, this dictionary was considered the oldest known compilation of Maltese vocabulary and formed the fundamental basis for numerous future dictionaries. *Ġamma* is a trilingual tridirectional dictionary that is divided into four volumes. Volume 1: Maltese-Italian-Latin; volumes 2 and 3: Italian-Latin-Maltese; and volume 4: Latin-Italian-Maltese. I limit myself to comment on Volume 1 which was published in 2016.

The manuscript dictionary consists of 1,036 folios. The text commences with a brief preface entitled “Lel Karrai” [To the reader], in which the compiler shares his motivation for compiling the dictionary. This is followed by a list of abbreviations. The dictionary begins with a collection of about 12,000 Maltese words (Brincat, 2011, p. 251), organised into 21 alphabetic stretches.

De Soldanis’s lexicographic method, as indicated by the title, has been descriptive from the outset. The lemmata he compiled were not selected according to etymological criteria, such as their origin from Carthaginian, but rather because they were current in the period. In “Lel Karrai,” the author states: “... believe me, reader, one can find foreign words, never the amount people anticipated, and when I recognised them, I uncovered them.”³ De Soldanis made the deliberate decision to use the Latin alphabet and exclude Arabic letters. This option was aimed to assist the potential users, as the few educated Maltese at that time were more accustomed to the Latin script than the Arabic one.

A notable deficiency in *Ġamma* is the lemmatisation of verbs in the first singular person of the imperfect, following the pattern found in dictionaries of Romance languages, rather than the third singular person of the perfect, which is the canonical form in many Arabic dictionaries due to the root-and-pattern morphology. Aquilina (1987, p. xxv) criticises the semantic treatment in *Ġamma* for being “generally prolix”. For example, one of the longest entries is “Gherk General” [Fungus Rock⁴], consisting of

³ My translation.

⁴ The name of a very small islet close to Dwejra in San Lawrenz, Gozo (Malta).

959 words. This copiousness is due to the fact that the lexicographer includes a wealth of encyclopaedic information which covers various topics such as food, traditional clothing, family habits, festivities, legends, proverbs, and many more.

The year 1796 holds great importance in the Maltese lexicographic tradition as it is the year in which *Ktýb yl Klým Malti 'Mfysser byl-Latġn u byt-Talġan* [*The Book of Maltese words explained in Latin and in Italian*, commonly referred to as *Lexicon*] was published in Rome. This dictionary was authored by Mikiel Anton Vassalli (Michaelis Antonii Vassalli), known as the father of the Maltese language, and it was the first printed dictionary of the Maltese language. Vassalli, an enlightened patriot and reformer, recognised that Maltese was mostly spoken by the illiterate masses, however he was the very first to call it “la lingua nazionale” [the national language]. He was not only the first to lay the foundations of the scientific study of Maltese but also devised a plan to develop and utilise it as a means of education and progress. The dictionary was therefore the third important publication for the implementation of this plan after an alphabet (1790) and a grammar (1791).

Vassalli's *Lexicon* is an alphabetical dictionary including around 18,000 lemmata (Vassalli, 1796, p. xxxviii). The book consists of 387 pages, with the central list comprising 341 pages, and two additional sections, namely the errata and ommissa. Despite being mentioned in the title page and in the actual text, the etymological and comparative appendices, along with the two indices (one in Latin and the other in Italian), were not published and are now lost (Bonavia, 2006). Vassalli commences the 44 pages front matter by introducing his dictionary to Maltese citizens with the address titled “Alla Nazione Maltese” [To the Maltese Nation]. Next, he proceeds to the “Discorso Preliminare” [Preliminary Discourse], which is organised into nine sections. Section 8 (“Of the present Maltese dictionary”) is particularly pertinent since it covers Vassalli's considerations and lexicographic principles, some of which are still relevant in the present day and hence, demonstrate to what extent this early work was based on scientific foundations.

In this section Vassalli (1796, p. xxxvii) states that the words for the dictionary were collected by using the index card system, a method previously employed by the renowned English lexicographer Samuel Johnson prior to 1755. Vassalli acknowledges the challenge of finding equivalents and states that when he couldn't find them, he provided paraphrases instead. Vassalli informs readers that he gathered information on various topics by consulting Maltese and Italian informants regarding words that were commonly used (p. xxxvii). Additionally, the author mentions that for the words he collected but didn't understand their meaning, he marked them with the abbreviation “d. s.” (i.e., “desire for meaning”). The *Lexicon* distinguishes between the primary or original meaning of a lemma and other meanings, both current and obsolete, which are indicated by Latin numbers. At the very end of the dictionary article, users are advised to consult the etymological appendix. The reason Vassalli gives for this decision is to avoid confusing those who are unfamiliar with Semitic languages.

The macrostructure is strict-alphabetical and although it was Vassalli first to recognise the root-and-pattern structure of the Semitic part of Maltese morphology, lemmata

were not grouped according to their roots. Instead, the derivatives are arranged according to their alphabetical order (p. xlv). The page is divided into two columns, with each column being assigned a number and displaying the initial two or three letters of the lemma signs it contains. These fulfil the identical purpose as the guide words to which we are nowadays used to in printed dictionaries. The Latin alphabet letters are listed on the left side of each column, starting at A and progressing in accordance with the lemmata in the column. These were meant to serve as a reference to the etymological appendix. Despite all this attention to the user's needs, Vassalli was often criticised because the alphabet he devised, although more precise than his predecessor's, contained "queer signs" (Saydon, 1953, p. 21) that were "unpardonably awkward" (Aquilina, 1994, p. 68). In addition, the alphabetical order does not correspond to that of the contemporary Maltese alphabet (Cachia, 2016, p. xxiv).

Vassalli's microstructure contains data types such as inflections (e.g., verb conjugation in the third person masculine in the imperfect) and Latin and Italian equivalents. Roman numerals were employed to separate senses. Entries are rather brief, typically ranging from 3 to 10 lines (Brincat, 2011, p. 257). In addition, Vassalli employs some symbols such as the equals sign (=) to indicate synonyms, and the cross symbol (✕) to represent "dictio Barbara" (p. xlv), i.e., non-Semitic words.⁵

Vassalli's *Lexicon* captured the interest of the few scholars who wrote about aspects of Maltese lexicography. Saydon (1953, p. 20), Cassar Pullicino (2001, pp. 68–69), and other researchers note that despite having consulted *Ġamma* (p. xviii), Vassalli excluded a significant number of Romance words that had been previously registered by De Soldanis. At the same instant, Vassalli collects most Maltese words of Semitic provenance. This exposes Vassalli's prescriptive stance against "barbarous expressions."⁶ Saydon (1953, p. 21), notes "omissions of meanings and idioms," while Aquilina (1994, p. 68) also observes "insufficiency of phraseology, idioms and prepositional verbs which expand the area of meanings."

Critics have also identified additional shortcomings. Vella (2015, p. 278) and Saydon (p. 21) contend that Vassalli deliberately selected a specific dialectal variety from among five regional dialects, and was subsequently described in the dictionary. This was the variety of the central area of *Hal Qormi* and *Haż-Żebbuġ*, as it sounded closest to Chaldean⁷ (Vella, 2015, pp. 284–289). Saydon (1953, p. 21), and Cachia (2016, p. xxv) find some shortcomings in the equivalents although Cachia believes that they are generally precise. According to Saydon (1953, p. 21), and Cachia (2016, p. xxv), Vassalli included several archaic words or even words of Semitic origin that were never used in Maltese. The most comprehensive analysis of this aspect can be found in Vella (2015, pp. 309–315). Vella convincingly argues that Vassalli has attempted to revive documented old Maltese words, such as "rajjes" [captain or commander] and

⁵ Inconsistencies in this respect are noted by Vella (2015, pp. 314–315).

⁶ Expressions imported in Maltese from Neo-Latin languages, mainly Italian. His position played a significant role in establishing the myth of the "Malti safi" [pure Maltese] which still finds some support even today.

⁷ Another term for Old Aramaic which is related to Phoenician: the language which Vassalli believed Maltese originated from. Phoenician is an extinct Canaanite Semitic language originally spoken in the region surrounding the cities of Tyre and Sidon (nowadays part of Lebanon).

“talamid” [student] (p. 309). Vella also notes that Vassalli has: created words derived from Semitic languages or roots that have no evidence in the available documents, such as “tynkit” [punctuation] (p. 311); introduced new calques, such as “Ktȳb yl Klȳm” [Wordbook] (p. 311); and even proposed inflections, like “vjerrejn” (dual) and “vjerrijȳt” (plural) for “vjerr” [journey] (singular) (p. 311).

Brincat (2011, p. 257) considers Vassalli’s *Lexicon* to be “significantly superior to its predecessors” while Saydon (1953, p. 20), commends its exceptional attributes of completeness, accuracy, and reliability, and asserts that until 1953 it remained the primary dictionary for Maltese, even with the publication of more recent dictionaries.

2.2 Bilingual and Trilingual Dictionaries Featuring English

In 1800, Malta became a British colony. This led to the introduction of English in Maltese dictionaries. In the 19th century, there were several dictionaries published by lexicographers who were followers of Vassalli, thus influenced by his works. Among these lexicographers there are Francesco Vella (1843); Giovanni Battista Falzon (1845; 1882) and Vincenzo Azzopardi (1847; 1849; 1853; 1856) (see Aquilina, 1987, pp. xxv-xxvi). All of them compiled trilingual dictionaries which, having been somewhat influenced by Vassalli’s *Lexicon*, critics have not given them much attention. Due to limitations in length, I will refrain from providing further commentary on these dictionaries and focus on the bilingual dictionaries. Nevertheless, I believe they merit further study.

In 1885, Salvatore Mamo published the first bilingual English-Maltese Dictionary. Mamo utilises the alphabet that was established by the Xirka Xemija [Semitic Association] in 1882, which is very similar to the alphabet that eventually gained official recognition (Privitera, 1964, p. 27). The *English-Maltese Dictionary* has a one-page introduction, a list of ten abbreviations, and a 260-page alphabetical central list, followed by two pages of errata corrige. Aquilina (1987, p. xxv) asserts that the dictionary gives “practically word for word explanations. It is more a vocabulary than a dictionary.” In order to maintain the dictionary’s “smallest possible dimensions, consistent with its completeness as a dictionary,” Mamo tells users in the preface that only the basic lemma is provided whereas derivations were excluded because he regarded them as “quite useless” for users who are already familiar with the basic lemma.

Yet, he did find room for both obsolete lemmata – many of which are indicated with (o) – and loanwords that are typographically marked in italics and provided as a last option in the entry, even though in the preface to a later (lost) dictionary, Mamo acknowledges that there is “unquestionable preference in colloquial intercourse to the pure Maltese words” (Ellul, 2023, p. 101). Although Saydon (1953, p. 23), criticises Mamo for his omission of idioms and other expressions as well as his failure to distinguish between equivalents, among other shortcomings, he ultimately concludes that “On the whole the book is fairly complete and reliable.”

In 1900, two dictionaries authored by Vincenzo Busuttil were published. These dictionaries can be regarded as a bidirectional bilingual Maltese-English and English-

Maltese dictionary. They are titled *Diziunariu mill Malti għall Inglis* [Dictionary from Maltese to English] (2 ed. 1932) u *Diziunariu mill Inglis għall Malti* [Dictionary from English to Maltese]. In both dictionaries, Busuttil used the Maltese alphabet of the Accademia Filologica. The English-Maltese dictionary consists of a one-page preface, followed by a central list of 1,389 pages. This list comprises roughly 90,000 English lemmata, arranged in alphabetical order. The back matter includes a list of 4 pages featuring initialisms and abbreviations along with their English and Maltese meanings (e.g., C.T. – Certified Teacher, Surmastru li għandu ċ-ċertifikat jew il-warrant; N.P. – Notary Public, nutar) as well as a page entitled “Lil min Jakra” [To the Reader] calling on readers to inform the compiler of any shortcomings they might find in his work.

Busuttil’s Maltese-English dictionary contains a page entitled “Lil min Jakra” [To the Reader] followed by a central list of 723 pages, which includes around 30,000 Maltese words distributed in alphabetical order. Not taking into account the cross-references would lead to a significant decrease in the number of lemmata. The back matter consists of 34 pages listing initialisms and abbreviations along with their definitions in Maltese and English, followed by 10 pages listing personal names with their corresponding meanings.

Agius (1990, p. 174) describes Busuttil’s dictionaries as “The first more comprehensive bilingual dictionaries [...] these contain a considerable number of scientific terms of Romance origin.” According to Privitera (1964, p. 27) and Saydon (1953, p. 23) Busuttil’s Maltese-English dictionary’s greatest merit is its comprehensive nature, as it includes both Semitic and Romance words without any etymological bias. Nevertheless, Aquilina (1987, p. xxv) observes “a deplorable reluctance to use loan-words for Maltese words of scientific or technical meaning borrowed from Italian or English.” Busuttil attempts to expedite the user’s searches by providing different realisations of the same lemma in different grammatical forms. As regards verbs, for example, he gives the imperative (e.g., “aġħmel” [do]), the canonical form in the third person masculine singular (e.g., “għamel” [to do/he did]) and the second person singular (e.g., “taġħmel” [you do]). With regards to nouns, he provides the singular (e.g., “ċens” [emphyteusis]; “żiemel” [horse]) and the broken plural forms accompanied by a cross-reference to the lemma in the singular (e.g., “ċnus” [emphyteuses]; “żwiemel” [horses]). Busuttil does not provide grammatical description of the lemmata, but he deserves credit for gathering several Maltese words and, particularly, expressions that have greatly assisted subsequent compilers. These are beneficial not just for users seeking a deeper comprehension of them in English or for those who want to utilise them in Maltese, but most crucially, they uncover changes in meaning that depend on the manner in which the lemma is used in expressions.

Monseigneur Carmelo Psaila’s⁸ *Dizzjunarju Ingliz u Malti* [English and Maltese Dictionary] was published in fascicles between 1936 and 1955. It has a 5-page introduction that exposes the compiler’s slightly prescriptive stance. This is followed by a comprehensive central list of 1,473 pages.

⁸ Malta’s national poet.

Psaila's dictionary has captured the interest of numerous critics, who have predominantly expressed positive feedback. For instance, Aquilina (1987, p. xxv-xxvi) argues that this dictionary is "an improvement on that of Busuttil," making it "the most comprehensive bilingual dictionary" of his time. Psaila's dictionary is widely acknowledged by critics to have been based on *The Concise Oxford English Dictionary of Current English* (1934). Psaila is mostly denounced by critics because according to them he resorted to paraphrasing instead of finding equivalent forms, as well as for excluding numerous Maltese words of non-Semitic origin from the dictionary.

In an upcoming study, Ellul discovered significant *Oxford* influences, specifically on the central list, the semantic treatment of entries, and even the layout of this dictionary. Regarding equivalence, Ellul discovered instances where Psaila unnecessarily employed paraphrasing, despite the fact that, contrary to common portrayal, the dictionary frequently offers cognitive or translational equivalents (e.g., meteoric-meteoriku, television-televiżjoni, visa-viża) that greatly aid users in their productive purposes. Ellul notes that a number of technical terms are missing in Psaila's dictionary. Furthermore, it seems that in certain instances, Psaila deliberately chose to omit these terms. However, it is incorrect to claim that Psaila was hesitant to incorporate specific terminology just because it is not Semitic. It is evident that the lexicographer made an attempt to enhance the Maltese vocabulary in order to keep it abreast with the progress made by mankind in the first half of the 20th century. Ellul regards Psaila's dictionary as a very important source for the introduction of new vocabulary that was subsequently collected in later dictionaries, notably *Il-Miklem Malti* by Erin Serracino-Inglott and the *Maltese-English Dictionary* by Joseph Aquilina. Ellul discovered that 8.8% of the lemmata in Aquilina's dictionary were extracted from Psaila's.

In 1987, Joseph Aquilina started publishing the most prestigious bilingual dictionary in the Maltese lexicographic tradition. Aquilina's is a bidirectional bilingual comprehensive dictionary that consists of six volumes: a two-volume *Maltese-English Dictionary* (1987–1990) and a four-volume *English-Maltese Dictionary* (published posthumously in 1999–2000). This review will focus on the ***Maltese-English Dictionary***.

Besides the central list of 41,016 lemmata (Brincat, 2011, p. 407), the dictionary's megastructure is comprised of both front matter and back matter featuring both integrated and non-integrated outer texts.

The dictionary's macrostructure presents an interesting compromise devised by the lexicographer in order to deal with two different – frequently interwoven – morphologies: root-and-pattern and stem. The extensive use of cross-references provides a strict-alphabetical entry as the start of the search route in a dictionary consulting procedure. This in turn leads the user to nests containing either strict-alphabetical ordering subject to textual condensations (in case of concatenative morphology) or morphologically ordered lemmata (in case of root-and-pattern morphology) with their relative treatment.

The dictionary's microstructure varies according to the lemma's part of speech and its derivative morphology, among other variables. However, some general features are consistently present throughout the whole dictionary. Users are directed into the dictionary article through a lemma sign which is typographically marked in bold capital letters. In case of stem-based lemmata, the stem is demarked by a vertical line. Occasionally, very limited pronunciation aid is given with the use of accents marking stress mainly to highlight the difference between homographs such as "*kaħħal*" [to plaster] and "*kaħħâl*" [plasterer]. Lemmata are accompanied by orthographic variants.⁹ Both lemmata and its variants are typically followed by an abbreviation "of the first or one of the earliest" Aquilina (1987, p. xiv) sources in which they were recorded, generally previous dictionaries or word-lists. The comment on form section comprises of roots, grammatical description and inflections such as dual and plural suffixes in case of nouns and, with regards to verbs, the passive participle and the vowels needed to form the imperfect.

The right-core subsection is marked with sense numbers and other typographical indicators such as italics and brackets. The basic equivalent in English is given first followed by a bracket containing usage examples and their translation. Whenever needed, this is preceded by stylistic or usage data indicating, for example, technical or figurative usage. The equivalent is then followed – without any structural indicators – by collocates, idioms, proverbs etc as well as their renditions in English. No distributional data is given as such, except for the record of the region where the word was attested, and this is placed in square brackets right after the sense number. After the treatment of all derivatives and compounds populating the nest, etymological data is found within square brackets at the very end of the microstructure.

2.3 Monolingual Dictionaries

The only published¹⁰ comprehensive monolingual dictionary of Maltese is Erin Serracino-Inglott's *Il-Miklem Malti* [*The Maltese Wordbook*] which was published between 1975 and 1989. In 2003, Mario Serracino-Inglott – the lexicographer's son – released the tenth volume of the dictionary which is made up of an Addendum and a Root Index. The Dictionary collects over 34,000 lemmata in 3,354 pages (including the Addendum).

The megastructure of *Il-Miklem Malti* comprises of front matter (mainly in the first volume), the strict alphabetically ordered central list and the Addendum (the 10th and last volume). Among the front matter items, the user can find the "Tagħrif u Tifsir" [Information and explanation] giving information about what to expect in the dictionary and usage guidelines.

⁹ Besides real orthographic variants, there is a considerable number of "variants" that are actually incorrect versions of the lemma sign, which were mainly recorded in previous dictionaries when the Maltese alphabet and orthography were not standardised.

¹⁰ The first significant endeavour to compile a comprehensive monolingual dictionary for Maltese was undertaken by Mikiel (Michael) Agius, who submitted his manuscript to the University of Malta's Library in 1978. Agius also compiled the manuscript of the first Maltese thesaurus. For detailed reviews of Agius's dictionary and thesaurus, read Ellul 2018a, 2018b, and 2022.

The most important part of *Il-Miklem Malti*'s back matter is a comprehensive root index. This index gathers all lemmata derived from the same root, which are scattered across the dictionary due to its rigorous alphabetical arrangement; hence giving also the lemma's page numbers. For instance, for the root K-T-B the compiler gives "kitba" [writing] p. 116; "kiteb" [to write/he wrote] p. 116 ... "mikteb" [office; desk] p. 142 and "nkiteb" [was written] p. 309. This is particularly beneficial for those who have an interest in studying the Semitic component in the Maltese language and its development.

As pointed out by Agius (1990, p. 179) this dictionary is a historical one in the sense that the compiler documents the source that registered the lemma before him – if it was not collected by himself in the first place – and gives the etymology of every lemma, whenever possible. However, in doing this, Serracino-Inglott opted to start the dictionary article with an abbreviation of the source, hence relegating the lemma signs (typographically marked in bold capital letters) as only second in line. The comment on form section comprises of grammatical description and inflections such as dual and plural suffixes in case of nouns and, with regards to verbs, the third person singular of the imperfect.

The right-core subsection starts off with the etymology given between brackets and typographically marked in bold. It is usually concise; however, examples of extensive etymological treatment are also abundant and these may make it very difficult for the users who are not interested in etymology to locate other data they need from the dictionary article.

The semantic comment starts, when applicable, with usage labels indicating figurative or technical usage such as *t.rel* (religious term). The bracketed sense numbers follow suit, introducing the senses (first the original sense, then current senses: basic, secondary and figurative) definitions, synonyms, phrases, idioms and proverbs marked in bold, and italicised examples mark the end of the microstructure.

Il-Miklem Malti has been criticised with regards to a number of issues, but mainly for the spelling of some words (e.g., Aquilina, 1977; Z. M., 1975); for including words of Italian origin which are at best uncommon or maybe hardly ever used at all; for excluding some English words which are common in Maltese such as *aircraft carrier*, *bullet*, *bumper*, *best man* etc; and for omitting the current English plural suffixes in favour of prescribing Maltese sound plural suffixes such as in *bejbijiet* (babies) and *bajsikilijiet* (bicycles) instead of *bejbis* and *bajsikils* respectively.

Albeit having these defects, the importance of *Il-Miklem Malti* in our lexicographic tradition can never be underestimated, since to date it is the only comprehensive monolingual dictionary based on scientific principles. As Agius (1990, p. 179) puts it:

"This dictionary has unquestionably important contributions to make to comparative Arabic dialectology. But it also has contributions to make to comparative Romance dialectology. The monolingual character of the lexicon should not prevent the non-Maltese researcher from consulting it. Indeed, its scientific approach should assist linguists to explore further their own particular fields of enquiry."

3. Recent Developments

Several lexicographic works published in the past 24 years owe much to Joseph Aquilina's and Erin Serracino-Inglott's comprehensive dictionaries. Amongst these, there are various editions of the monolingual *Id-Dizzjunarju Malti u Teżawru ta' Malti Mhaddem* [*The Maltese Dictionary and thesaurus of current Maltese*] by Mario Serracino-Inglott (2000; 2003; 2011; 2016); the *Concise Maltese-English/English-Maltese Dictionary* by Joseph Aquilina (2006); *The English-Maltese Dictionary for the 21st Century* by Charles Briffa (2015); *It-Teżawru* [*The Thesaurus*] by Albert W. Agius (2000; 2010) and *Kelma għal Kelma: Teżawru Tematiku Malti* [*Word for Word: Maltese Thematic Thesaurus*] both in print and online by Dwayne Ellul (2020). Additionally, there have been initial attempts to create online dictionaries. Two notable examples were the *Malta Online Dictionary* by Carlo Saliba and *il-miklem.com* by John Spiteri Gingell. Given the word count constraint, this section focuses on two initiatives associated with the University of Malta: John J. Camilleri's *Ġabra* [*Collection*] accessed from the Institute of Linguistics' website, and *Ġabra Terminologika* [*Collection of Terms*] by the Department of Maltese at the University of Malta.

3.1 Ġabra

Ġabra [*Collection*] is a free open lexicon for Maltese described by Camilleri (2016, p. 376) as “an online database of lexical resources with full inflection forms” which have been automatically generated using the Grammatical Framework developed by Ranta (2011). Ġabra is based on the idea of integrating together various existing lexical resources into a single database.¹¹ These resources include: a list of 4,142 root-and-pattern verbs; a corpus of 654 broken plurals for nouns and adjectives; a list of over 2,500 verbal nouns; two smaller separate lexicons; and a basic morphological generator for Maltese from Apertium.

Although Ġabra is not a complete dictionary for Maltese, it includes 20,382 entries and 4,514,683 inflectional word forms. Several of these are linked by root, include English equivalents and are marked for various morphological and syntactic features, such as verbal patterns and transitivity. According to the lexicon's web page:

“The level of detail [within the entries] varies, and there are words, forms, and senses which are missing. Many of the word forms in the database are automatically generated, and may contain errors.”

Camilleri (2013, p. 377) stresses the need for Ġabra to be a full-form lexicon by arguing that traditional dictionaries – which only provide the lemma and affixes that “cannot be blindly appended to the headword” – do not offer sufficient information to the users and that they would require some further knowledge of the language in order to construct the inflection.

¹¹ Some database snapshots can be downloaded from <https://mlrs.research.um.edu.mt/resources/gabra-api/p/download>

As stated by Camilleri (2013, p. 379), the *Ġabra* lexical resources have varying structures. To ensure adaptability to future resources, the database needs to be “flexible enough to support heterogeneous data as it is found.” Subsequently, he proceeds to outline the top-level structure for lexemes, word forms and roots. Camilleri (p. 380) informs us that the database engine he chose for implementing *Ġabra* is MongoDB.

Ġabra can be accessed on <https://mlrs.research.um.edu.mt/resources/gabra/>. Results can be obtained by utilising three distinct search methods: the default mode, advanced search, and root search. A single search yields outcomes from the complete database, encompassing all lemmas, inflected word forms, and English glosses. When the user clicks on the lemma in the results page, they are redirected to another page that has a significantly larger amount of data: a phonetic transcription of the lemma in IPA, a description of the part of speech, a list of senses or equivalents, the root, morphological and syntactic features, a list of related items, and an acknowledgement of the source. Additionally, a list of possible inflections in the perfect, imperfect, and imperative is provided, along with the option to include the suffixes for the direct and indirect objects. The polarity can be changed to enable access to both positive and negative inflectional forms.

Ġabra includes a system that allows users to report any errors or omissions they may encounter.

3.2 *Ġabra Terminoloġika*

Over the past years, the Department of Maltese at the University of Malta has conducted dissertations to compile terminology from many registers, including both traditional activities (such as carpentry, hunting and trapping, and fishing with feather bait) and non-traditional activities (such as hairdressing, clothing, and pasta). During this period, there was a growing demand for the development of the Maltese language in all its domains. This was particularly important for numerous EU translators who frequently had to translate legal texts and instructions containing technical terms from different sectors or trades (Portelli 2017).

The *Ġabra Terminoloġika* [*Collection of Terms*] project commenced in 2018 with the objective of merging the glossaries from dissertations (40 at the time of writing) into a unified database accessible to both researchers and the general public. This language planning approach to terminology is in line with the initial five of eight stages identified by Cabré (1999: 49) that should be followed in language modernisation.

The database is presently populated with around 4,000 published entries covering 14 specialised fields such as “bini” [construction]; “ekonomija” [economics]; “ilbies” [clothing]; “mastrudaxxa” [carpentry]; and “żwiemel” [horses]. Another 1,766 entries are also queued for editing before publication.

The *Ġabra Terminoloġika* database is a WordPress database that utilises the MySQL database management system. The entries may include many data types, which

are indicated by red data identifying indicators. These data types include lemma, lemma variations, English equivalent, grammatical description, inflections (such as feminine, dual, plural, collective), definition in Maltese, source, and field (register). Furthermore, the incorporation of images, movies, and audio is also possible. Every entry is assigned to one or more specialised fields.

In order to enhance user experience, the database can be filtered according to the initial letter or the specialised field. Users can utilise the search feature to find matches across all entries and their respective fields.

Ġabra Terminoloġika can be accessed on <https://malti.mt/termini/glossarji/>. Users have three options for searching: they can utilise the search box located in the upper right corner, go through the *oqsma* [fields], or browse according to the initial letter of the lemma by selecting the blue icons. There are also two outer features: *Riżorsi Terminoloġiċi* [terminological resources] and *Dwarna* [about us]. The terminological resources page is a collection of references mostly focused on print, audio, or audiovisual resources related to the flora and fauna found in *art u sema* [land and sky] and *baħar* [sea].

A search for a term (e.g., “fagu” [beech wood]) will display the following data: the lemma “fagu”; a communicative English equivalent¹² “beech wood”; grammatical description “nom maskili” [noun, masculine]; definition in Maltese (translated as: “hard and heavy pinkish timber with a reddish grain”); specialised field [“qasam”] and source [“ħajr”]. Furthermore, *Ġabra Terminoloġika* can be searched for the terms of a whole field, e.g., carpentry. The lemmata pertaining to the field are presented in alphabetical order, allowing the user to select the desired entry for viewing.

4. Conclusion

This study looked at the most important dictionaries in the Maltese lexicographic tradition, tracing their evolution over centuries. From an analysis of various aspects of the most important dictionaries, it appears that the Maltese historical and cultural environment was not only fundamental in the weaving of the Maltese language but also in the development of its dictionaries. During the Knights’ period, the dominant Latinate culture in Malta required the inclusion of Italian or Latin language in dictionaries, alongside Maltese. However, once Malta came under the British rule, these languages were gradually replaced by English.

An important observation is that, prior to the standardisation of the alphabet, lexicographers often created their own alphabets in order to effectively represent the sounds of the Maltese language or to improve previous alphabets in order to make it easier for users. An inclination towards favouring the description and use of the Semitic component of the Maltese language, while neglecting the more

¹² This is not always possible due to anisomorphism between Maltese and English, particularly in more traditional activities, as pointed out by Gómez González-Jover (228) amongst others. However other data in the microstructure may help “in the different decision-making processes of comprehension and production of texts.”

contemporary vocabulary derived from Romance or English origins, is also noted in certain dictionaries. The preferred macrostructure is often strict-alphabetical order. However, lexicographers such as Aquilina have developed an interesting compromise that involves nesting and cross-referencing.

Although the Maltese lexicographic tradition is still going strong, with dictionaries being compiled every now and then and with ongoing online projects such as *Ġabra* and *Ġabra Terminoloġika*, a comprehensive online Maltese dictionary still needs to be compiled.

References

Dictionaries mentioned

Agius De Soldanis, G. P. F. (2016). *Ġamma tal-Kliem Kartaginis Mscerred fel Fom tal Maltin u Ghaucin*. (R. Carabott, Ed.) Il-Kunsill Nazzjonali tal-Ilsien Malti, il-Librerija Nazzjonali, id-Dipartiment tal-Malti, l-Akkademja tal-Malti, Heritage Malta, it-trust tad-Dipartiment tal-Innovazzjoni fir-Riċerka u l-Iżvilupp tal-Università ta' Malta, BPC International u l-Ministeru ta' Ghawdex. Florence. (Original work finished c. 1767)

Aquilina, J. (1987–1990). *Maltese-English Dictionary*. Midsea Books.

Busuttil, V. (1900). *Dizjunariu mill Malti għall Inglis*. Malta.

Busuttil, V. (1900). *Dizjunariu mill Inglis għall Malti*. Malta.

Camilleri, J. J. (2013). *A Computational Grammar and Lexicon for Maltese*. [M. Sc Thesis, Chalmers University of Technology].

Cassola, A. (1992). *The Biblioteca Vallicelliana "Regole per la Lingua Maltese"*. Said International.

Dipartiment tal-Malti, l-Università ta' Malta. *Ġabra Terminoloġika*. <https://malti.mt/termini/glossarji/>

Mamo, S. (1885). *English-Maltese Dictionary*. Aquilina e Co. Retrieved July 20, 2024, https://books.google.com.mt/books?id=DcwNAAAAQAAJ&printsec=frontcover&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false

Psaila, C. (1936–1955). *Dizzjunarju Ingliz u Malti*. DOI.

Serracino-Ingloft, E. (1975–2003). *Il-Miklem Malti*. Midsea Books.

Vassalli, M. A. (1796). *Ktîb yġ Klîm Mâlti 'Mfysser byl-Latîn u byt-Talġân*. Rome. Retrieved July 20, 2024, from https://books.google.com.mt/books?id=uQpaAAAACAAJ&printsec=frontcover&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false

Other literature

Agius, D. A. (1990). Al-Miklem Malti: A contribution to Arabic lexical dialectology. *British Society for Middle Eastern Studies*, 17 (2), 171–180. Retrieved July 20, 2024, from <http://jstore.org/stable/194709>

Agius De Soldanis, G. P. F. (2007). *Della Lingua Punica Presentamente usata da Maltesi & C.* (J. Zammit Ciantar, Ed.) Facsimile Edition Malta. (Original work published in 1750)

Aquilina, J. (1977). L-Ortografija Llum: I: Fehmiet dwar l-Ortografija tal-Malti. *Il-Malti*, 53 (2–4) 3–16. Retrieved July 20, 2024, from <https://www.um.edu.mt/library/oar/bitstream/123456789/62027/1/Il-Malti%201977%2053%282-4%29.PDF>

Aquilina, J. (1994). Maltese Lexicography. In J. Aquilina (Ed.), *Papers in Maltese Linguistics* (pp. 63–74). Malta UP.

Bonavia, C. (2006, October 29). How Vassalli Lost His Library. *The Sunday Times of Malta*, 61. Retrieved July 20, 2024, from <https://timesofmalta.com/article/when-vassallis-library-was-torn-apart.37211>

Brincat, J. M. (2011). *Maltese and Other Languages: A Linguistic History of Malta*. Midsea Books.

Cabré, M. T. (1999). *Terminology: Theory, Methods and Applications*. John Benjamins Publishing Company.

Cachia, L. (Ed.). (2016). *Lexicon jew Dizzjunarju Malti-Latin-Taljan*. (An edition of Vassalli's *Lexicon* published in 1796)

Camilleri, J. J. (2016). Digitizing the Grammar and Vocabulary of Maltese. In G. Puech, & B. Saade (Eds.), *Shifts and Patterns in Maltese* (pp. 359–385). De Gruyter, Inc.

Cassar Pullicino, Ġ. (2001). *Il-Kitba bil-Malti sal-1870*. PIN Publications.

Cassola, A. (1986, October 19). The first printed list of Maltese words. *The Sunday Times of Malta*, 21.

Ellul, D. (2018). *Il-Kompilazzjoni ta' Teżawru Tematiku Malti dwar in-Natura tal-Bniedem u r-Relazzjonijiet Soċjali Tiegħu*. [Unpublished Master's thesis]. University of Malta.

Ellul, D. (2018). Harsa Hafifa lejn l-Ewwel Dizzjunarju Malti-Malti u Teżawru miġbura minn Michael Agius. *Lehen il-Malti*, 37, 104–120.

Ellul, D. (2022). Is-Sisien tad-“Dizzjunarju tal-Lsien Malti” ta' Michael Agius u Aspetti mill-Makrostruttura u l-Mikrostruttura Tiegħu. *Lehen il-Malti*, 41, 166–185.

Ellul, D. (2023). Sittax-il folja mid-Dizzjunarju Manuskritt Mitluf ta' Salvatore Mamo: storja tad-Dizzjunarju Manuskritt u Tqabbil mad-Dizzjunarju l-Iehor ta' Mamo Stampat fl-1885. *Lehen il-Malti*, 42, 83–103.

Gómez González-Jover, A. (2006). Meaning and Anisomorphism in Modern Lexicography. In M. L'Homme (Ed.), *Processing of Terms in Specialized Dictionaries: New Models and Techniques* (pp. 215–234). John Benjamins Publishing Company.

Mifsud, M. (1995). *Loan Verbs in Maltese: A Descriptive and Comparative Study*. E.J. Brill.

Portelli, S. (2007). Terminology, Bilingualism and Language Planning (or lack thereof): The

Case of Maltese. In P. Faini (Ed.), *Terminological Approaches in the European Context* (pp. 98–106). Cambridge Scholars Publishing.

Privitera, V. (1964). *Id-Dizzjunarji Maltin*. [Unpublished Bachelor's dissertation] University of Malta.

Ranta, A. (2011). *Grammatical Framework: Programming with Multilingual Grammars*. CSLI Publications.

Saydon, P. P. (1953). Bibliographical Aids to the Study of Maltese. Retrieved July 20, 2024, from *Journal of Near Eastern Studies*, 12 (1), 20–30. <https://www.jstor.org/stable/542741>

Svensén, B. (2009). *A Handbook of Lexicography: The Theory and Practice of Dictionary-Making*. Cambridge University Press.

Vella, O. (2015). *Il-Kodifikazzjoni tal-Ilsien Malti fis-Seklu 18 fl-Isfond Kulturali*. [Doctoral dissertation, University of Malta]. OAR@UM

Z. M. (1975) “Il-Miklem Malti” ta’ Erin Serracino-Inglott – L-Ewwel Volum – Klabb Kotba Maltin, 1975 St. Paul’s Press Ltd. *Il-Malti*, 51(4), 119–120. Retrieved July 20, 2024, from <https://www.um.edu.mt/library/oar/bitstream/123456789/61592/1/Il-Malti%201975%2051%284%29.PDF>

Acknowledgements

I express gratitude to the anonymous reviewers for their careful reading of my manuscript and their insightful comments and suggestions.

Contact information

Dwayne Ellul

University of Malta

dwayne.ellul@um.edu.mt

Ivana Filipović Petrović and Kristina Kocijan

CREATING THE DATASET OF CROATIAN VERBAL IDIOMS

Automatic Identification in a Corpus and Lexicographic Implementation

Abstract This research proposes a step forward in the automatic identification and analysis of verbal idioms in Croatian. The use of the NooJ automated text processing tool, along with the MaCoCu corpus and the *Online Dictionary of Croatian Idioms* (ODCI), provides a robust framework for recognizing and categorizing these multi-word expressions (MWEs). The research comprises two parts: (a) creation of a dataset by utilizing the ODCI that allowed for a set of 898 verbal idioms to be compiled and annotated with linguistic features, including structure, morphological features, and variation patterns; (b) analysis of extracted data that provides insights into the lexicographical and linguistic significance of the idioms, such as variability, modification, and frequency of use. The study highlights the challenges posed by idiomatic variations and the verb's role as the most variable component in idioms. For instance, the idiom *soliti pamet komu* ('to give unsolicited advice') is often modified for expressiveness, such as in the phrase "having a big saltshaker to salt everyone's mind." The dataset aims for lexicographic integration into ODCI and supports the creation of electronic language resources. It also contributes to theoretical and cross-lingual research, with the CLARIN repository expected to enhance data reusability in NLP. The study's findings offer a deeper understanding of verbal idioms' dynamics and their computational processing.

Keywords verbal idioms; automatic identification of multi-word expressions; linguistic features of idioms; NooJ; lexicographic analysis

1. Introduction

The computational processing of figurative language, particularly multi-word expressions (MWEs), is a significant area of interest in nowadays computational linguistics and phraseology. The goal is to create language resources and effectively use linguistic data in various applications such as natural language processing (NLP), machine translation, and large-scale language models (LLMs).

This has brought the following methodological issues to the fore: how to accurately describe the syntactic and semantic features of MWEs, given their complex structure, variability, and figurative meanings, the set of features well known as a pain in the neck of NLP (Sag et al., 2002). In other words, how to formulate correct search criteria, computationally process the data, and finally, use the results for practical applications like lexicography.

In our research, we focus on addressing these issues by introducing a method for the automatic identification of sentences containing verbal idioms in Croatian, with the aim of creating a lexicon of Croatian verbal idioms.¹

Like many other low to middle-resourced languages, Croatian has lacked digital phraseological resources until the recent publication of the *Online Dictionary of the Croatian Idioms*² (ODCI) version 1.0 freely accessible since 2022. This dictionary is corpus-driven and has been processed by linguists and lexicographers, making it a reliable source for the initial description of verbal idioms and their usage.

Previous research in this field (Pasquer et al., 2020a; 2020b) has shown that verbal idioms pose a challenge as they occur in different variants, such as word order or lexico-syntactic forms. Corpus studies on Croatian verbal idioms (Filipović Petrović & Parizoska, 2017; Parizoska, 2022) have indicated that the verb is the most varied component, while the other components are relatively fixed, e.g., *gurati/pomesti/staviti pod tepih* ‘to hide a problem’ (lit. push/sweep/put sth under the carpet=N Acc sg). Some idioms also have open slots that can be filled with various adjectives, adding to their complexity, e.g., *poprimiti* ADJ *razmjere* which translates to ‘reach ADJ proportions’ (epidemic/dramatic/unprecedented).

To tackle these challenges, we have utilized the NooJ tool (Silberztein, 2016) and automatic identification (Ramisch et al., 2020; Gantar & Krek, 2022) for the second methodological issue. NooJ is a rule-based automated text processing tool that supports the detection of MWEs in running texts. Previous studies (Machonis, 2012; Kocijan & Librenjak, 2016; Najar et al., 2017; Rhazi & Boulaalam, 2018) have validated the NooJ tool’s efficacy in identifying multi-word expressions, which is a testament to its utility in linguistic research. It combines a dictionary and grammar to identify elements that co-occur in specific positions. This approach allows for the identification of both known verbal idioms and potential new lexical and syntactic realizations thus facilitating us in extraction of MWEs from the large corpora for the purpose of lexicon building. For this specific task, either a list of MWEs from existing lexicons or manually annotated corpora are used as the basis, and the result is a set of corpus sentences containing MWEs in all typical syntactic and semantic realizations.³

Based on the described features that were implemented in NooJ, we automatically identified the sentences that include verbal idioms in Croatian, resulting in over 132,000 corresponding examples from the corpus. The final step involved linguistic evaluation of the results and performance of the automatic procedure.

¹ In this paper, we use the term *idiom*, which is equivalent to the term *phraseological unit* in the terminologies of some languages, and it refers to MWE in the narrowest sense. We rely on the division of MWEs from Kosem, Krek & Gantar (2020), according to which phraseological units are multiword lexical units with their own metaphoric meaning that speakers most often use when they want to express something in a noticeable and expressive way, differently from the neutral, for example *stealing days from God* instead of *wasting time*.

² <https://lexonomy.elex.is/#/frazeloskirjecnikhr>

³ The other method of automatic MWEs detection refers to the identification of MWEs in corpora regardless of existing MWEs. Both methods have already been researched, for example, for Slovenian (Škvorc, Gantar & Šikonja, 2022; Gantar & Krek, 2022).

In addition to describing the method that can be used for research in the field of computational phraseology, the intention of this study was also to produce a dataset of automatically identified Croatian verbal idioms, which, as a freely available linguistic resource, can be used for various purposes in the field of NLP, theoretical phraseology and cross-lingual research. For this paper, we utilized the obtained dataset for lexicographical purposes, specifically to supplement the existing ODCI with data from a more contemporary corpus.

The paper is structured as follows. The introductory section is followed by Section 2 which outlines the methodology for identifying verbal idioms within the corpus, including the initial list creation, formal description of linguistic features, design of the dictionary and grammar in NooJ, details of the corpus, and the extraction method employed. Section 3 discusses the evaluation of the obtained data. In Section 4, we delve into the lexicographic analysis and implementation of the dataset, particularly its use in enhancing existing phraseological dictionaries. Finally, in Section 5, we conclude and discuss implications for future research.

2. Methodology of Verbal Idioms Identification in the Corpus

The method used for the automatic identification of verbal idioms in the corpus comprises four steps. First, an initial list of verbal idioms is compiled from freely available phraseological resource for the Croatian language (Filipović Petrović & Parizoska, 2022) which serves as a starting point for the description and classification of morpho-syntactic models of verbal idioms. This step includes the preparation of the data for the second step – the creation of the dictionary and grammar in the NooJ tool. The default settings in NooJ are then used to search the corpus and identify phraseological expressions. The final step involves manually evaluating the results to create a dataset of verbal idioms.

The chosen method has several advantages. Firstly, the ODCI is freely available which is an important prerequisite for using the data. Also, it is regularly updated, so the results of this research can be practically applied to this dictionary. Additionally, previous studies have demonstrated the effectiveness of the NooJ tool for identifying multi-word expressions in Croatian text by combining information provided at the level of a dictionary with algorithms developed over syntactic grammar (Kocijan & Librenjak, 2016; 2018). Moreover, a new resource, the CLASSLA web corpus has become available in 2023, providing a large and diverse collection of contemporary Croatian texts suitable for phraseological research.

2.1 The Initial List

To create the initial list of Croatian verbal idioms, we used data from ODCI. It is accessible through the Lexonomy Elexis platform and based on hrWaC, the most extensive Croatian corpus at the time of its initial release (Ljubešić & Klubička, 2016). This dictionary was created using a corpus-driven method, where lexicographers collected idioms by extracting the thousand most frequent verbs in the hrWaC corpus using the Sketch Engine tool. Then, using collocation, filter, and word sketch functions, they extracted conventionalized

multi-word expressions with figurative meanings (Filipović Petrović & Parizoska, 2022). In the extraction of verbal idioms, we adhered to the morpho-syntactic and semantic criteria (Fink-Arsovski, 2002). Namely, if the meaning of the idiom is expressed by a verb or a group of verbs, it is a verbal idiom (e.g., ‘to manage, control, operate’ as in *vući konce* ‘pull the strings’). Also, when incorporated into discourse, verbal idioms are in the function of a predicate. In this way, we utilized a set of 898 verbal idioms for the initial list.

2.2 Formal Description of Linguistic Features

To create a NooJ dictionary and grammar, it was necessary to formulate the linguistic features of idioms. The list of 898 idioms was analyzed and relevant linguistic characteristics were attached to each idiom. Based on these characteristics, verbal idioms were categorized according to their grammatical structure into several basic models and submodels, as exemplified in Table 1⁴.

Table 1: Excerpt of the classification of verbal idioms by grammatical structure into models and submodels

Model	Extended submodel	Idiom
V + N		<i>raditi frku</i> (‘make a fuss about sth’) <i>puniti kesu</i> (‘line your pockets’)
	+ PREP with N	<i>nositi gaće na štapu</i> (‘on the breadline’) <i>tjerati mak na konac</i> (‘pigheaded’)
V + PREP with N		<i>ispasti iz kolotečine</i> (‘to get off track’) <i>kipjeti od bijesa</i> (‘to boil with rage’)
	+ REC	<i>opaliti po džepu koga</i> (‘hit smb’s pocketbook’) <i>nasmijati se u brk komu</i> (‘laugh in smb’s face’)
V + KAO (‘AS’) + N		<i>spavati kao klada</i> (‘sleep like a top’) <i>piti kao smuk</i> (‘drink like a fish’)
	+ PREP with N	<i>nicati kao gljive poslije kiše</i> (‘to mushroom’) <i>snalaziti se kao riba u vodi</i> (‘take to sth like a duck to water’)

Each component of the idiom is assigned additional grammatical description, as shown in Table 2.⁵

Table 2: Example of models with additional grammatical description

Model	Idiom
TRANS V, all tense, pers. + N Acc pl	<i>vući konce</i> ‘pull the strings’
NEG V form, all tense, pers. + N Acc sg + PREP with N Instr sg	<i>ne vidjeti prst pred nosom</i> ‘one can’t see a thing’

⁴ In Table 1, the following symbols are used: V – verb; N – noun; PREP – preposition; REC – rection (a grammatical relationship where a verb or preposition determines the case of a noun or noun phrase).

⁵ In Table 2, the following symbols are used: TRANS V – transitive verb; NEG V form – negative verb form; all tense, pers. – all tenses and persons; N – noun; Acc – accusative case; Instr – instrumental case; pl – plural; sg – singular; PREP – preposition.

Some idioms are completely fixed, with the verb appearing in a single person and tense form, as in *nije o glavu* ('it's not a hanging matter') and *nigdje ne gori* ('what's the rush?'). In such cases, the components of idioms are entered into the NooJ dictionary as uninflected words. The NooJ dictionary also contains all elements of an idiom, while the grammar defines the order of idiom components. This is because the verb does not necessarily need to be the first component in the expression. For example, in the idiom *od govna napraviti pitu* ('make a silk purse out of a sow's ear'), the verb is positioned in the middle.

2.3 Dictionary and Grammar Design in the NooJ

Two main engines in NooJ are the dictionary and grammar. The dictionary determines the elements that occur together, and grammar defines the position of each element. The grammar complements the dictionary and the two work in unison.

A special feature, denoted as +FXC (Frozen Expression Components), is employed to aid in distinguishing between individual word occurrences and their idiomatic usages within multiword expressions (Kocijan & Librenjak, 2016). For instance, the verb *objesiti* (meaning 'to hang') is a transitive verb and when used as in sentence (1), we aim to treat it as a cohesive unit. To accomplish this, we will introduce an entry (2) in the NooJ dictionary. However, in sentence (3), the focus is not on analyzing *objesiti* as a standalone verb but rather on its idiomatic usage. Consequently, we will supplement the dictionary with an additional entry for this verb, augmented with an explanation of its idiomatic structure (4).

- 1) *Ana je objesila rublje.* 'Ana hung the laundry outside to dry.'
- 2) *objesiti*,V+FLX=UGASITI
- 3) *Ana je objesila nos.* 'Ana became dejected' (lit. hung her nose)
- 4) *objesiti*,V+FXC+FLX=UGASITI+Model=V_N+N=<nos,N+Case=Acc>

When the noun *nos* (meaning 'nose') in the accusative case and the verb *objesiti* ('to hang') appear together in a sentence, NooJ prioritizes the annotation from the second dictionary entry (4).

The rule variation is specified through the +**Model** attribute assigned to each dictionary entry. This attribute clarifies potential ambiguity when the same verb is used in various idioms, such as *pobratu batine* ('to receive a beating') compared to *pobratu vrhnje* ('reap the benefits'). In the same vein, the verb *ići* ('to go') forms part of idiomatic expressions such as *ići čijim stopama* ('to follow in someone's footsteps') shown in example (5), and *ići preko leševa* (literally, 'to walk over corpses') which translates to 'stop at nothing', illustrated in example (6).

- 5) *ići*,V+FXC+FLX=IĆI+**Model**=V_XN+N=<stopa,N+I+p>
- 6) *ići*,V+FXC+FLX=IĆI+**Model**=V_PP+S=preko+NP=<leš,N+G+p>

In the dictionary, we specify which elements can coexist, and in grammar design, we determine the placement of each element and their broader context when necessary. For example, the algorithm depicted in Figure 1 is designed for two patterns:

1. Verb (V) + Noun (N) [upper branch] and
2. Verb (V) + intermediary element (X) + Noun (N) [lower branch].

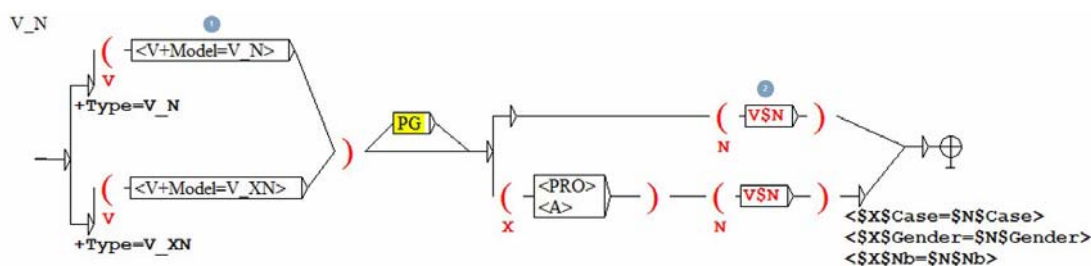


Fig. 1: Algorithm that recognizes models V_N and V_XN

The first pattern aligns with example (3) and connects to dictionary entry (4). The algorithm moves from the leftmost side to the rightmost node. It starts by pinpointing the verb defined in the dictionary with the attribute Model=V_N (Figure 1, marker 1). This verb is then stored as variable \$V. The next node (Figure 1, marker 2) uses this variable to locate the noun that pairs with that specific verb, distinguishing it from other verbs in the dictionary. Thus, if the verb *objesiti* ('to hang') is followed by the noun *nos* ('nose'), the algorithm recognizes both words as an idiomatic pair, rather than as separate entities.

Furthermore, Croatian grammar's nuances allow for an auxiliary verb between the main verb and noun in idiomatic expressions, especially when forming complex verb tenses (as in example 3). An optional node [PG] is included to account for this.

The same algorithm applies to the pattern Verb (V) + intermediary element (X) + Noun (N), as seen in *ići čijim stopama* ('follow in someone's footsteps') and *poprimiti epidemijske razmjere* ('reach epidemic proportions'). The grammar's lower branch (Figure 1) describes idiomatic structures with the attribute Model=V_XN, permitting a pronoun or adjective before the noun, stored as variable \$X. In examples like *poprimiti epidemijske razmjere* ('reach epidemic proportions'), the noun must match the preceding variable \$X in case, gender, and number as defined in the last node of the lower branch.

The ODCI corpus data underscored the prevalence of inversion in Croatian idioms, necessitating its inclusion in our description. For instance, *od govna napraviti pitu* ('turn sth worthless into sth of value') commonly appears as PP V N in the hrWaC corpus, with variations like *od govna pitu napraviti* (PP N V), *napraviti od govna pitu* (V PP N), and *napraviti pitu od govna* (V N PP). Despite multiple possible versions, the dictionary contains only one entry for all the listed variations of the idiom (7).

- 7) napraviti,V+FXC+FLX=CIJEPI+Model=V_NP_
PP+N1=<pita,N+Acc>+S=od+NP=govna.

The entry (7) consists of the verb *napraviti* ('to make'), the accusative noun *pitu* ('a pie'), and the prepositional phrase *od govna* ('from shit'), which can appear in positions 1, 2, and 3 (Figure 2). Although all components are listed in the dictionary, their possible arrangements are defined through the grammar (Figure 2).

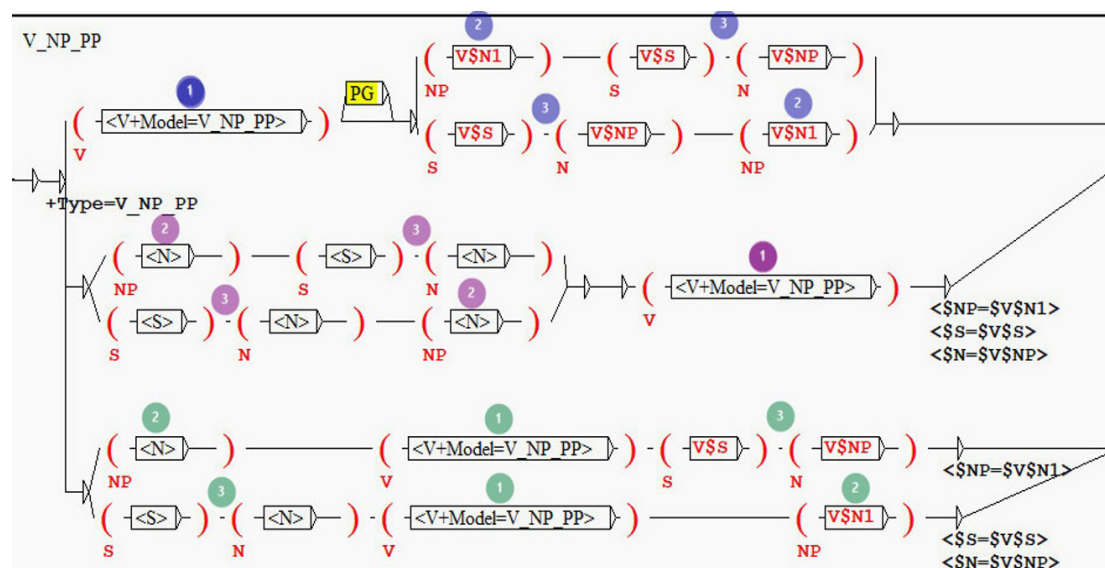


Fig. 2: Excerpt of an Augmented algorithm for V + N + Prepositional Phrase idiomatic structure and its inversions

2.4 Corpus

The CLASSLA web corpus for Croatian stands as one of the seven newly established corpora for South Slavic languages at the CLASSLA Knowledge Centre of the Slovenian CLARIN consortium. Originating from the MaCoCu project (Bañón et al., 2022), web crawling techniques were employed to amass monolingual and parallel datasets, as described in Ljubešić & Kuzman (2024). The corpus's resency, volume, and the text variety render it exceptionally apt for phraseological studies. The final CLASSLA-web corpus for Croatian is hosted on the CLARIN.SI repository, labelled as the Croatian CLASSLA-web.hr corpus (Ljubešić, Rupnik & Kuzman, 2024), encompassing 2,575 million tokens across 5,422 thousand documents.

2.5 Extraction Method

Based on our input described in sections 2.2. and 2.3., NooJ's data processing identified over 132,000 examples that correspond to corpus sentences. The unpredictability of the file sizes necessitated sophisticated technical support. Specifically, the data extraction, which took more than 49,000 working hours across four servers, occasionally exceeded the resources available to us.

Furthermore, the data obtained indicated the need for additional manual cleaning and removal. This was due to the method of corpus collection and the nature of the source data, which were web pages. Therefore, the results included various types of data. These included machine-generated text such as postdates and coded repetitions, as well as constructions like “learn more”, “read more”, “sent from my...”. The results also contained sentences truncated into multiple lines and significant text repetition characteristic for forums, especially when authors referred to previous writings.

Considering all these factors, along with the time constraints of our conference paper submission deadline, we decided to present the results for the hundred most frequent verbs, using approximately 80% of the corpus data. We will continue with data extraction and analysis until we create a final dataset to be a stand-alone resource in open access.

3. Data Evaluation

When 70% of the files were processed, we compiled a list of the 100 most frequently occurring verbs in the results. This resulted in a list of 50,987 lines that we manually evaluated linguistically. Figure 3 shows the top 10 most frequent models, with the model „V S N“ at the top, occurring 12,920 times.

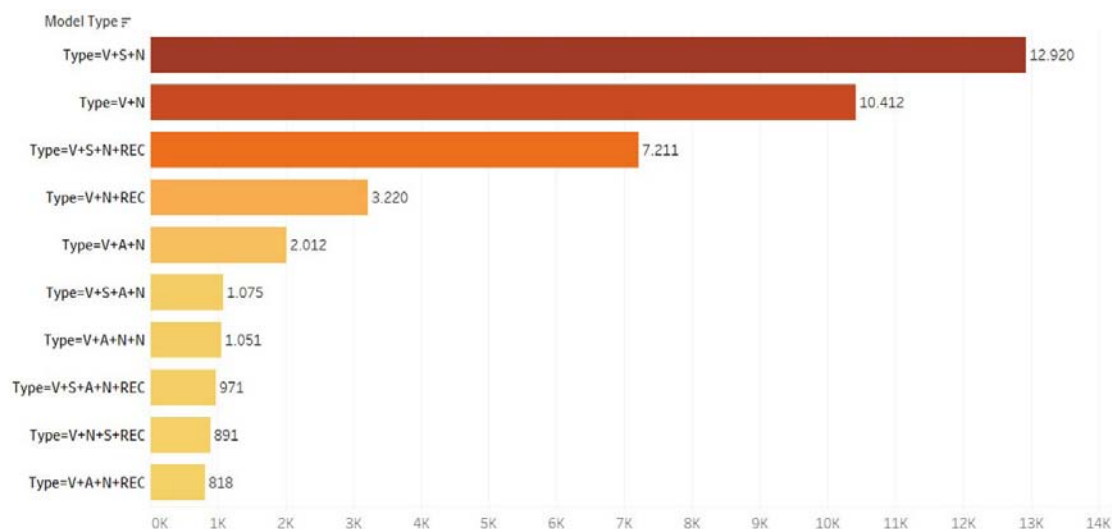


Fig. 3: Distribution of idioms within the top 10 model types

Among these verbs, *biti* (‘to be’), *imati* (‘to have’), *poći* (in this specific context, meaning ‘to get’), *postaviti* (‘to set’), and *ići* (‘to go’) yielded the most idiomatic examples overall. Verb *biti* with a total of 38 idioms, has the greatest variety of idiomatic usage, ranging from only one occurrence (*biti kao bubregu u loju* lit. to be like a kidney in fat ‘to be in clover’) up to 655 occurrences (*biti na izmaku* lit. to be nearing the end ‘to be running out of steam’). On the other hand, *poći* is detected only in one idiom, *poći za rukom* (‘to manage to do something’), but it occurs 2,466 times, making it the most frequent idiom in use within our corpus so far. It is followed by 1,816 occurrences of *postaviti temelje* (‘to lay foundations’), 1,222 occurrences of *pasti u zaborav* (‘to be

forgotten'), 1,051 occurrences of *imati pune ruke posla* ('to have hands full of work' meaning 'to be very busy'), and 1,040 occurrences of *uzeti maha* ('to gain ground'). The majority of the manual evaluation involved removing instances that contained the desired components but formed examples of literal usage in the language. For example, we obtained 1073 instances of the idiom *isplivati na površinu* ('to come to light, to surface'). Out of these, in 703 instances, it appeared as an idiom, meaning 'to appear, suddenly become visible, noticeable' as in example (8), while in 370 instances, it had the literal meaning 'to emerge from a liquid' (as in example 9).

- 8) *Nije trebalo dugo da **na površinu isplivaju** dobre i loše strane takvog načina rada.* 'It didn't take long for the good and bad sides of such a working method to come to light.'
- 9) *Njoke pažljivo stavite u posudu i kuhajte par minuta dok ne **isplivaju na površinu**.* 'Carefully place the gnocchi in a pot and cook for a few minutes until they float to the surface.'

On the other hand, the construction *postaviti temelje* ('to lay foundations') appeared in 8 literal contexts out of a total of 1824 appearances. Table 3 shows which idioms are used both with their idiomatic and literal meaning ranging from 40% up to 60% of times. The size of the square before the percentage corresponds to the number of occurrences in the corpus. For example, *biti na nogama* (lit. to be on one's feet) 'to be up and about' and *dati crveni karton* (lit. to give a red card) 'to eject' are both used in 53% of times as an idiom. However, the total number of occurrences of *biti na nogama* in the corpus is 873 which is much greater compared to the number of occurrences of *dati crveni karton* which appears only 32 times.

Table 3: Excerpt of multiword expressions that appear in their idiomatic meaning between 40% and 60%

IDIOM	idiomatic meaning		literal meaning	
dignuti na stražnje noge	•	60,00%	•	40,00%
izbijati na površinu	■	59,27%	■	40,73%
izaći na površinu	■	58,50%	■	41,50%
izlaziti na površinu	■	56,89%	■	43,11%
stati na noge	■	55,59%	■	44,41%
baciti na ulicu	•	55,10%	•	44,90%
biti na nogama	■	53,49%	■	46,51%
dati crveni karton	•	53,12%	•	46,88%
naći na ulici	■	52,49%	■	47,51%
biti u sjeni	■	52,37%	■	47,63%
bacati mrvice	•	50,00%	•	50,00%
baciti u komu	•	50,00%	•	50,00%
biti u prvim redovima	■	46,98%	■	53,02%
servirati na tanjuru	•	40,00%	•	60,00%
baciti na cestu	•	40,00%	•	60,00%
ići na nos	•	40,00%	•	60,00%
držati u sjeni	•	40,00%	•	60,00%

4. Lexicographic Analysis and Implementation

In this section, we analyze extracted data from a lexicographic perspective. To describe a language in a dictionary, a lexicographer needs to gather linguistic data, specifically data on how the language is used. Idioms, which are multi-word expressions with figurative meanings, pose a challenge not only for NLP but also for lexicographic processing. Our research has shown that automated procedures for identifying idioms in a corpus can help lexicographers by providing comprehensive lists of sentences with idiomatic usage and statistical data that can be extracted from them. The data revealed insights about verbal idioms in terms of linguistically significant aspects such as variability, modification, and frequency of use. These insights have a direct impact on lexicographically substantial categories such as macrostructure, entry structure, usage notes and examples of use.

4.1 Macrostructure and Entry Structure

As presented in Section 3, we gathered data on the frequency of verbal idioms in Croatian, identifying the most common ones and the most productive ones. Three verbs seem to dominate in the productivity domain, namely verbs *biti* ('to be') with 38 idioms, *imati* ('to have') with 23 idioms and *ići* ('to go') with 19 idioms (Figure 4).

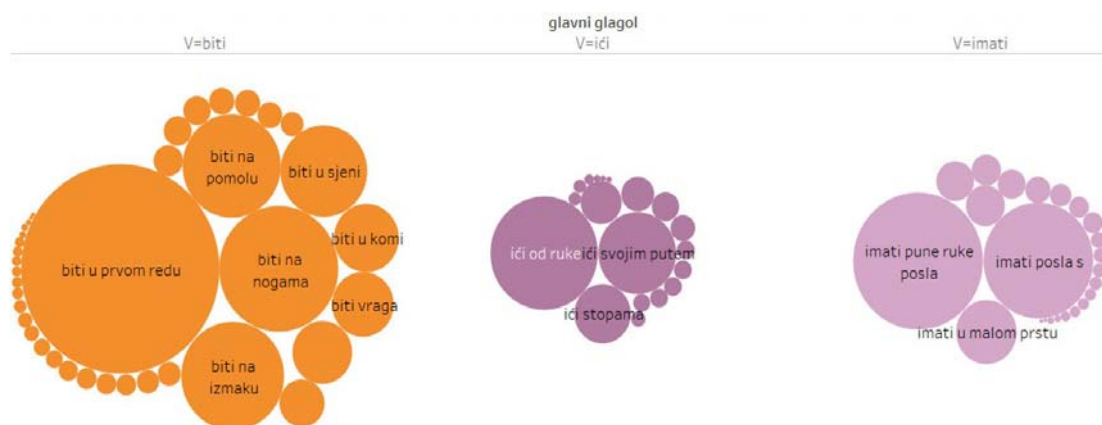


Fig. 4: Distribution of the top 3 most productive verbs: *biti*, *imati*, *ići*

Further analysis revealed that for idioms with a variable verbal component, the frequency of occurrence can serve as a guide for selecting the canonical form and determining the lexical variants to be included in the entry. For instance, construction *na svojoj koži* ('on one's skin'), mostly detected with the idiom *osjetiti na svojoj koži* ('to feel on one's skin') appeared with the verb *osjetiti* ('to feel') in 472 occurrences in the corpus, and with the verb *iskusiti* ('to experience') in 34 occurrences. Component that occurs with the largest variety of verbs is *na površinu* (literally meaning 'on the surface') which in most instances appears with the verb *izaći* ('to come out') (703 occurrences) and in somewhat smaller number of occurrences with verbs *izbijati*, *izbiti*, *izlaziti*, *izvući* and *izbaciti* as shown in Figure 5.

Based on the frequency criterion, which is used for selecting the canonical form of the idiom in ODCI, the form *izaći na površinu* should occupy the position of the headword. Depending on the lexicographic decision, two other variants that are most common after *izaći* can be included as variants in the entry.

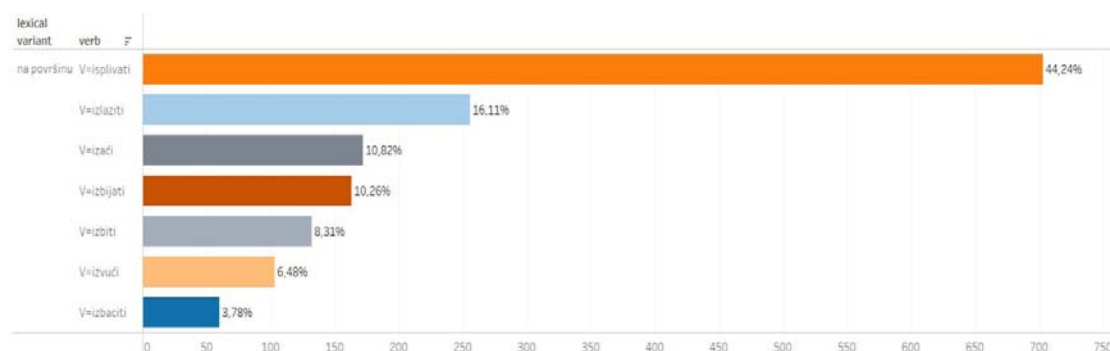


Fig. 5: Diversity of verbs occurring with component 'na površinu'

4.2 Usage Notes

4.2.1 Negation

The data also showed a pattern of frequent use of certain verbal idioms in specific contexts. For example, the idiom *imati i ovce i novce* 'to have it all' (lit. 'to have both sheep and money') often appears with a negation: *you cannot have both sheep and money*. This type of data, known as usage notes, has undergone significant development in lexicography thanks to corpus data upon which dictionaries are based, as large corpora provide evidence about contexts in which idioms are typically used. Usage notes particularly come into focus in phraseological dictionaries, taking into account the importance of context for understanding the meaning of idioms. Additionally, they are highly valuable for language learners. A significant number of verbal idioms appear with negation in corpus, some of which are more commonly used with negation than in the affirmative form, such as *poći za rukom* 'to manage to do something', with only 3,75% appearances in the affirmative form. Usage notes may be grammatical, such as the data on negation, and sociolinguistic (see Rundell, 2023), such as data on intentional *ad hoc* modifications of idioms in the speech act, which we discuss in the following paragraph.

4.2.2 Idiom Modifications

A comprehensive list of sentences has also revealed certain patterns in the use of variants and modifications of idioms. This is yet another phenomenon of language in use that can only be documented in a corpus-based dictionary (cf. Langlotz, 2006), and this research has shown that automatic identification is a method that significantly enhances the detection of intentional alterations of idioms in language. During the linguistic evaluation of the results, we observed several modified uses of the idiom *soliti pamet* 'to advise someone in a condescending way' (lit. to salt someone's mind)

in sentences from the corpus. Speakers deliberately change idiom for expressiveness by playing with the literal meaning of the idiom components, as in the example (10), or expand it with new components semantically close to the literal use of the idiom's components, as in the example (11).

- 10) *Gradonačelnik je bio u pravu kad je kazao kako su to ljudi koji imaju jednu veliku soljenku i vole svima soliti pamet.* 'The mayor was right when he said that these are people who have a big saltshaker and love to salt everyone's mind.'
- 11) *Gdje bi nam bio kraj da nemamo vas da nam solite pamet i biberite rane?* 'Where would we be without you to salt our mind and pepper our wounds?'

5. Conclusion

In this study, we tackled the complex challenges related to the computational processing of figurative language, particularly MWEs, by focusing on Croatian verbal idioms. Our primary goal was to develop a reliable method for the automatic identification of these idioms, thereby creating a valuable linguistic resource for various applications in NLP, machine translation, and large language models (LLMs). By doing so, we hope to significantly enhance the sophistication and contextual awareness of language technologies and enable more accurate translations, nuanced sentiment analysis, and more natural interactions in chatbots and virtual assistants, among other applications.

Utilizing the NooJ tool, we successfully identified over 132,000 sentences containing Croatian verbal idioms from a large corpus. This significant achievement was made possible through a combination of dictionary-based and grammar-based approaches. These methods enabled us to capture both confirmed and potentially novel lexical and syntactic realizations of idioms.

The resulting dataset can enhance the existing Online Dictionary of Croatian Idioms (ODCI) by providing detailed information on the canonical forms of idioms and their most frequent variant forms. Additionally, it offers insights into specific usage patterns that are relevant to dictionary users and highlights intentional modifications made by speakers during communication. Finally, the rules and methodologies established for Croatian verbal idioms in this study have the potential to be adapted for use in other Slavic languages, thereby broadening the scope and applicability of our findings.

References

Bañón, M., Esplà-Gomis, M., Forcada, M. L., García-Romero, C., Kuzman, T., Ljubešić, N., van Noord, R., Pla Sempere, L., Ramírez-Sánchez, G., Rupnik, P., Suchomel, V., Toral, A., van der Werff, T., & Zaragoza, J. (2022). *Croatian web corpus MaCoCu-hr 1.0*. Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, <http://hdl.handle.net/11356/1516>

Filipović Petrović, I., & Parizoska, J. (2017). Leksikografska obrada frazema s promjenjivom glagolskom sastavnicom u hrvatskome. *Jezikoslovlje*, 18(2), 245–278.

- Filipović Petrović, I., & Parizoska, J. (2022). *Frazeološki rječnik hrvatskoga jezika*. Hrvatska akademija znanosti i umjetnosti. <https://lexonomy.elex.is/#/frazeoloskirjecnikhr>
- Fink-Arsovski, Ž. (2002). *Poredbena frazeologija: pogled izvana i iznutra*. FF press.
- Gantar, P., & Krek, S. (2022). Creating the lexicon of multi-word expressions for Slovene. Methodology and structure. In A. Klosa-Kückelhaus, S. Engelberg, Ch. Möhrs, & P. Storjohann (Eds.), *Dictionaries and Society. Proceedings of the XX EURALEX International Congress* (pp. 549–562). IDS-Verlag.
- Kocijan, K., & Librenjak, S. (2016). Recognizing verb-based Croatian idiomatic MWUs. In T. Okrut, Y. Hetsevich, M. Silberstein, & H. Stanislavenskapp (Eds.), *Automatic Processing of Natural-Language Electronic Texts with NooJ* (pp. 96–106). Springer.
- Kocijan, K., & Librenjak, S. (2018). The quest for Croatian idioms as multiword units. In R. Mitkov, J. Monti, G. Corpas Pastor, & V. Seretan (Eds.), *Multiword units in machine translation and translation technology* (pp. 202–221). John Benjamins.
- Kosem, I., Krek, S., & Gantar, P. (2020). Defining collocation for Slovenian lexical resources. In I. Kosem, & P. Gantar (Eds.), *Kolokacije v leksikografiji: Collocations in lexicography: existing solutions and future challenges*, Letn. 8, št. 2 (pp. 1–27). Znanstvena založba Filozofske fakultete, Slovenščina 2.0. <https://revije.ff.uni-lj.si/slovenscina2/article/view/9338/9069>, DOI: 10.4312/slo2.0. 2020.2.1-27
- Langlotz, A. (2006). *Idiomatic creativity: a cognitive-linguistic model of idiom-representation and idiom-variation in English*. John Benjamins.
- Ljubešić, N., & Klubička, F. (2016). *Croatian web corpus hrWaC 2.1*. Slovenian language resource repository CLARIN.SI, ISSN 2820-4042. <http://hdl.handle.net/11356/1064>
- Ljubešić, N., & Kuzman, T. (2024). *CLASSLA-web: Comparable Web Corpora of South Slavic Languages Enriched with Linguistic and Genre Annotation*. <https://arxiv.org/abs/2403.12721>
- Ljubešić, N., Rupnik, P., & Kuzman, T. (2024). *Croatian web corpus CLASSLA-web.hr 1.0*. Slovenian language resource repository CLARIN.SI.
- Machonis, P. (2012). Sorting NooJ out to take Multiword Expressions into account. In K. Vučković, B. Bekavac, & M. Silberstein (Eds.), *Formalising Natural Languages with NooJ: Selected Papers from the NooJ 2011 International Conference* (pp. 152–165). Cambridge Scholars Publishing.
- Najar, D., Mesfar, S., & Ben Ghezela, H. (2017). Inflectional and Morphological Variation of Arabic Multi-Word Expressions. In T. Okrut, Y. Hetsevich, M. Silberstein, & H. Stanislavenskapp (Eds.), *Automatic Processing of Natural Language Electronic Texts with NooJ: Selected Papers from the International Conference NooJ 2016* (pp. 37–47). Springer CCIS Series #667.
- Parizoska, J. (2022). *Frazeologija i kognitivna lingvistika*. Srednja Europa.
- Pasquer, C., Savary, A., Antoine, J.-Y., Ramisch, C., Labroche, N., & Giacometti, A. (2020a). *To Be or Not To Be a Verbal Multiword Expression: A Quest for Discriminating Features*. Retrieved May 10, 2024, from arXiv:2007.11381

Pasquer, C., Savary, A., Ramisch, C., & Antoine, J.-Y. (2020b). Verbal multiword expression identification: Do we need a sledgehammer to crack a nut? In D. Scott, N. Bel, & C. Zong (Eds.), *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 3333–3345). International Committee on Computational Linguistics.

Ramisch, C., Savary, A., Guillaume, B., Waszczuk, J., Candito, M., Vaidya, A., Barbu Mititelu, V., Bhatia, A., Iñurrieta, U., Giouli, V., Güngör, T., Jiang, M., Lichte, T., Liebeskind, Ch., Monti, J., Ramisch, R., Stymne, S., Walsh, A., & Xu, H. (2020). Edition 1.2 of the PARSEME shared task on semi-supervised identification of verbal multiword expressions. In S. Markantonatou et al. (Eds.), *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons* (pp. 107–118). Association for Computational Linguistics.

Rhazi, A., & Boulaalam, A. (2018). Corpus-based Extraction and Translation of Arabic Multi-Words Expressions (MWEs). In S. Mbarki, M. Mourchid, & M. Silberstein (Eds.), *Formalising Natural Languages with NooJ 2017 and its Natural Language Processing Applications* (pp. 143–155). Springer CCIS Series #811.

Rundell, M. (2023). Automating the creation of dictionaries: are we nearly there? In *Lexicography, Artificial Intelligence, And Dictionary Users. Proceedings of the 16th International Conference of the Asian Association for Lexicography: Lexicography* (Asialex 2023 Proceedings) (pp. 1–9). Yonsei University.

Sag, I., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. (2002). Multiword Expressions: a Pain in the Neck for NLP. In A. Gelbukh (Ed.), *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics* (CICLing-2002) (pp. 1–15). Springer-Verlag.

Silberstein, M. (2016). *Formalizing Natural Languages: The NooJ Approach*. Wiley-ISTE.

Škvorc, T. Polona, G., & Robnik-Šikonja, M. (2022). MICE: mining idioms with contextual embeddings. *Knowledge-based systems*, 235, 1–11. <https://doi.org/10.1016/j.knosys.2021.107606>

Contact information

Ivana Filipović Petrović
Croatian Academy of Sciences
Linguistic Research Institute
ifilipovic@hazu.hr

Kristina Kocijan
University of Zagreb
Faculty of Humanities and Social Sciences
Department of Information and Communication Sciences
krkocijan@ffzg.unizg.hr

Louise Holmer, Ann Lillieström, Emma Sköldberg, and
Jonatan Uppström

TIME TO SAY GOODBYE REVISITED

On the Exclusion of Headwords from the Swedish Academy Glossary (SAOL)

Abstract In the revision process of dictionaries, adding new headwords or new senses to already existing headwords is what typically receives the most attention. In this article, we bring into focus the intriguing dilemma of exclusion of headwords from the Swedish Academy Glossary (SAOL), which is still published in print versions. In the e-dictionary-era, removing headwords may seem questionable. SAOL is, however, a contemporary dictionary which aims to reflect present-day Swedish. In order to keep the lemma list up to date, new headwords are added and obsolete words are removed. The editors of SAOL have practised lemma exclusion in connection with the revisions of new editions for almost 150 years. In this paper, we present SAOL and argue that lemma exclusion is crucial to SAOL's aim and target group. We also present our most recent corpus material, methods and tools included in this process.

Keywords lemma exclusion; SAOL; Swedish monolingual dictionaries; Språkbanken Text

1. Introduction

When revising dictionaries, the revision of the lemma list and the inclusion of new headwords in the dictionary naturally gain a lot of attention. The corpus material, the methods, and the motivation behind the selection of headwords (lemmas), are also a topic for various investigations.

However, the exclusion of headwords from dictionaries is not as often a topic of discussion in metalexigraphy. One reason for this may be the difficulty in detecting words that are falling out of use – especially in comparison to newly coined words. The declining use of a word seems to be a slower process compared to the formation and spread of a new word. In addition, a word that has previously shown a tendency to fall out of use may return with renewed force (which seems to be the case for the Swedish verb *prokrastinera* ‘procrastinate’). Hence, as Gellerstam (2005, p. 133) points out, the exclusion of lemmas is a delicate matter in practical lexicographical work (cf. Diamond, 2016). Gellerstam also states that, as a result, obsolete words often remain in dictionaries as “silent witnesses from another time,” possibly with a usage label such as “uncommon”.

Nowadays, the possible deletion of lemmas might also be considered irrelevant. For e-dictionaries, space restrictions do not apply in the same manner as they do, or previously did, for print dictionaries.

The aim of the present study is to analyse the list of headwords in the Swedish Academy Glossary (SAOL, 14th ed., 2015) and to investigate how (in)frequently inflectional forms of each headword occur in a large, contemporary text corpus. The motivation behind this is that low frequency in present-day texts is the main reason for exclusion of lemmas from SAOL.

The paper is structured as follows: first, we give a brief overview of the dictionary in focus, SAOL, with special attention to its list of lemmas and the principles for inclusion and exclusion of headwords. In section 3, we present the methodology and material in our study. In section 4, the results of the investigations are presented and discussed, and finally, in section 5, we relate the results to the lexicographical challenges concerning space restrictions in a dictionary that is published in print and in various e-versions.

2. The List of Headwords in SAOL

SAOL plays an important role in Swedish linguistic infrastructure (Holmer et al., 2024). It is a monolingual dictionary of present-day Swedish, focusing on orthography and inflection, but not semantics or phraseology. It is aimed at supporting production-related tasks and may also provide information on word formation such as linking morphemes in solid compounds (Berg & Malmgren, 2010). Moreover, SAOL is often regarded as an unofficial norm for spelling (Holmer et al., 2015; Sköldberg et al., 2019) and it is being widely used in schools, by teachers, in university courses, by linguists, by writers and by officials writing public documents.

There are several other Germanic languages with dictionaries similar to SAOL, for example, Duden's *Die deutsche Rechtschreibung* for German, *Woordenlijst Nederlandse Taal* for Dutch, *Retskrivningsordbogen* for Danish, *Tanums store rettskrivningsordbok* for Norwegian (Bokmål) and *Stafsetningarorðabókin* for Icelandic (see Gellerstam, 2009b; Vikør, 2009; Berg et al., 2010). They all have a tradition of regular revisions, and they were all published in print first, and have later, in various ways, been developed into different e-versions. See Figure 1 for an example of the print version of SAOL 14 (2015), with lemmas starting with the letter R.

re.kond.it.ion.era v. ~de ~t • åter försätta i god kondition re.kond.it.ion.er.ing s. ~en ~ar re.kon.stru.era v. ~de ~t • återställa i ursprunglig form; ombilda; jfr rekonstruktion re.kon.strukt.ion s. ~en ~er • till rekonstruera re.kon.strukt.ions plan s. ~en ~er	re.kord jämn adj. ~t ~a • till jämn 2 re.kord kall adj. ~t ~a re.kord lopp s. ~et; pl. ~ re.kord låg adj. ~t ~a • r. ränta re.kord lång adj. ~t ~a re.kord många pron.; pl. re.kord mäss.ig adj. ~t ~a re.kord nivå s. ~n ~er re.kord not.er.ing s. ~en ~ar re.kord prest.at.ion s. ~en ~er	rekryt [rekry't] s. ~en ~er • soldat under grund- utbildning rekryt.era v. ~de ~t • skaffa nya medlemmar el. pers- onal; nyanställa rekryt.er.are [-e'r-] s. ~n; pl. ~, best. pl. rekryterarna rekryt.er.ing s. ~en ~ar rekryt.er.ings bas s. ~en ~er rekryt.er.ings be.hov s. ~et; pl. ~

Fig. 1: Example from SAOL 14 (2015) in print, the letter R

As seen in Figure 1, SAOL lists lemmas, parts of speech and inflections in abbreviated forms for each headword, and in some cases short semantic descriptions. Solid compounds ending with a polysemous constituent, such as *jämn* in *rekordjämn* (lit. ‘record even’) (in the middle column), are followed by a reference to the relevant sense of the constituent (here *jämn* 2). The lemma *rekonstruktionsplan* (‘reconstruction plan’) (in the first column) illustrates the linking morpheme -s- between the constituents *rekonstruktion* and *plan*.

SAOL has about 126,000 lemmas distributed over the following major parts of speech:

- nouns: 94,000
- adjectives: 16,500
- verbs: 10,500
- adverbs: 1,200
- others: 3,800

Neologisms in Swedish tend to be nouns, often in the form of solid compounds. As a natural effect, candidate words for inclusion in SAOL are, therefore, often solid compound nouns; this also applies to candidates for exclusion (Berg et al., 2010).

SAOL has a long tradition in Sweden – its first edition was published in 1874 and it has since been fully revised and published in 14 editions over 150 years. From the first edition’s approximate 34,000 headwords to the latest edition and its 126,000 headwords, many new words have been included in the dictionary, but a fair share of obsolete ones have also been excluded (Holmer et al., 2024).

2.1 Lemma Inclusion in SAOL

As a rule, lemma inclusion criteria in SAOL involve frequency in written texts, a widespread use geographically, and preferably, use in texts that are aimed at the general public and non-specialists. Traditionally, the Swedish corpus material has mostly consisted of newspaper texts, but over time, the editors have also started to include other text types as well, e.g., social media (although edited texts are more relied upon than non-edited texts). See e.g., Gellerstam (2009a) for an in-depth study of the lemma lists in different editions of SAOL, and Berg & Malmgren (2010) for a discussion focusing on the inclusion of solid compounds in the glossary.

There are various aspects of lemma inclusion, and it is, in this study, impossible to give a satisfying overview of this field of research. Some (of many) recent studies include Klosa & Lungen (2018) on the methods of semi-automatic detection of new German words and their possible inclusion in the *Neologismenwörterbuch*. Another example is ten Hacken (2020) who discusses the relationship between new lemmas, the empirical reality, and the norms in general. Moreover, Sánchez & Maroto (2021)

provide a comparison between a more theoretical approach and speakers' insights in relation to new Spanish words.

When new editions of SAOL have been released (about every 10th year) the number of included words such as neologisms has received a considerable amount of attention in the media and among users. When it became known that the (then) forthcoming edition of SAOL 14 was going to include the gender-neutral pronoun *hen* in 2014, even international media wrote articles on the topic.

2.2 Lemma Exclusion in SAOL

A previous study of SAOL, named “Time to say goodbye”, focused on the exclusion of mainly solid compounds (Berg et al., 2010). It showed that about 10% of the headwords of SAOL 13 (2006, comprising about 125,000 headwords) were not in use in the reference corpus of about 250 million words (at that time considered a large corpus) of mainly newspaper texts – in spite of the fact that about 5,500 lemmas had been removed in relation to the previous edition.

An earlier study (Gellerstam, 2005) also showed similar results; about 10% of the lemmas of SAOL 12 (1998, with about 120,000 headwords) were not in use in a corpus of about 90 million words. It is important to note that irrespective of the number of headwords included in SAOL, the glossary will still only include a sample of contemporary Swedish vocabulary, since the possibilities of coining new compounds and derivatives are almost infinite (Berg et al., 2010; cf. Hanks 2019).

In an article by Svensson (2019) in *Språktidningen*, a Swedish popular science magazine, the author lists and comments on 1,001 headwords in SAOL 14 that may risk being excluded in the next revision due to low frequency in texts. Svensson asks the readers to adopt and use their favourite of these obsolete words in order to save them from endangerment. This campaign in *Språktidningen* was well received in the media, and also led to considerably more e-mails from the general public to the editors of SAOL.

The next edition of SAOL is planned for late 2025, and, as already mentioned, it will be published not only digitally but also in print. There is an explicit request from both members of the Swedish Academy and the SAOL editorial staff that the different versions of the glossary should include the same lemmas. In order to keep the print version accessible and in one volume, it is necessary to delete obsolete words and senses, although the digital version allows an expanded lemma list. The deletion of lemmas is therefore a prerequisite for making room for newly established ones (cf. Rundell, 2015, p. 302). The editors must also face the challenges of turning a former print dictionary, that has been revised to meet the demands of the online versions, back (or maybe forward) to a classical print dictionary again.

Regardless of its form of publication, SAOL is a synchronic dictionary and not a mausoleum for words that are not in use. Considering future studies of the Swedish vocabulary, the list of headwords in SAOL should represent contemporary Swedish.

For users interested in words that are no longer in use, the online resource SAOLhist provides the lists of headwords for all previous editions of the glossary (see section 4 for examples).

3. Methods and Selected Corpora

Kilgarrieff (2013, p. 74) states that "building a headword list for a new dictionary (or revising one for an existing title) has never been an exact science" (see also Trap-Jensen, 2013, on lemma selection in relation to corpus content). However, from the SAOL perspective, it is important that the editorial work with both the inclusion of new lemmas and the exclusion of existing ones is scientifically based, as far as possible.

The current SAOL editors have access to both language technology-based methods as well as considerably larger corpora with a broader coverage of different texts, than was the case for previous SAOL lexicographers (see section 2.2). The annotated corpora are available through Språkbanken Text's word research platform Korp (Borin et al., 2012; see also Holmer et al., 2024, pp. 71–72). The corpora consist of newspaper texts, from about 2010 to 2021, from different parts of Sweden. In addition, it contains texts published by Swedish Television (SVT), the public service television company, from the same period, the Swedish version of Wikipedia, and texts from a poetry site with user-generated content. Språkbanken Text's text analysis tool Sparv (Hammarstedt et al., 2022) provides frequency lists of the lemma forms of all words occurring in the corpora. This makes it straightforward to look up how often a headword in the dictionary is used and to produce a list of exclusion candidates. The total number of tokens in the corpora is nearly 3 billion.

4. Results of the Study

Even though several thousand lemmas have been removed between the editions of SAOL, studies have shown that earlier editions of SAOL include relatively many words that are not used in modern texts (Gellerstam, 2005; Berg et al., 2010; see section 2.2).

A preliminary comparison between the lemma list in SAOL 14 and the words that appear in our selected corpora shows that approx. 121,000 of the 126,000 lemmas appear in at least one of their inflectional forms in the text materials.

However, the remaining lemmas, about 4,100, are not used in the texts. This corresponds to approximately 3.5% of the lemmas. The result can be compared with Berg et al. (2010) who show that approximately 10% of the lemmas in SAOL 13 were not used in their reference corpus. Still, it is important to remember that the corpus used in our study is more than 10 times larger, and that the text materials in our corpora represent different genres etc.

Figure 2 shows a selection of examples of lemmas which are not used in the current corpus in any inflectional form.

atomtekniker ('atomic engineer')	rövarromantik ('robber romance')
bofinksbo ('chaffinch nest')	sandalrem ('sandal strap')
dambekantskap ('lady acquaintance')	senvårdag ('late day of spring')
dolikocefali ('dolichocephaly')	sexmejla ('(to send) sex e-mail')
civiltjänstgörare ('person doing civil service')	sladdrare ('gossiper')
eau-de-vie-butelj ('bottle of eau-de-vie')	tevetelefon ('TV telephone')
ellipsformig ('formed like an ellips')	tändsticksetikett ('match(box) label')
fingerlik ('fingerlike')	vattenösa ('water scoop')
foderpotatis ('feeding potato')	väsensfrändskap ('being's kinship')
förtursbehandla ('treat with priority')	yawlriggad ('yawl rigged')
förhandlingspraxis ('negotiation practices')	ångerköpthet ('remorse')
grodmansutrustning ('frogman equipment')	änkeleken ('the widow game')
kakelugns-galler ('tiled stove grate')	ögonflörta ('(to) flirt with the eyes')

Fig 2: Examples of lemmas in SAOL 14 (2015) which are not used in any inflectional form in the corpus

Among the lemmas in Figure 2, there are nouns such as *dambekantskap* ('lady acquaintance'), *dolikocefali* ('dolichocephaly'), *foderpotatis* ('feeding potato'), adjectives such as *ellipsformig* ('formed like an ellips'), *fingerlik* ('fingerlike'), and verbs such as *förtursbehandla* ('treat with priority') and *ögonflörta* ('(to) flirt with the eyes'). Some of the words are registered in several SAOL editions, for example, *sladdrare* ('gossiper'), which has been included since the first edition from 1874. However, many lemmas, such as *fingerlik* ('finger like'), *rövarromantik* ('robber romance'), *tändsticksetikett* ('match(box) label') and *väsensfrändskap* ('being's kinship'), have been present in SAOL since the ninth edition (from 1950), which holds the largest list of lemmas to the present day (approx. 155,000 lemmas).

A more qualitative review of the entire list of so-called non-matches, i.e., SAOL lemmas that are not used in the current corpora, shows that these represent slightly different types of words. Many non-matches consist either of older designations that have been replaced by new ones (such as *hörselvårdsassistent* lit. 'hearing care assistant') or designations for objects, phenomena, activities, etc. no longer in use in the surrounding society (such as *tevetelefon* 'TV telephone' and *skråtobak* 'chewing tobacco').

Furthermore, in the list of non-matches, there are a few, simplex words (such as *pyngel* 'trifles of little value') and a considerable number of solid compounds (such as *rövarromantik* and *tevetelefon*). In particular, complex solid compounds like *grodmansutrustning* ('frogman equipment') and *kakelugns-galler* ('tiled stove grate') are frequent in the list.

According to Berg & Malmgren (2010), a lot of "intuitively slightly odd derivations" were excluded in the revisional process of SAOL 14. Despite this, there are still many derivatives among the non-matches. Among them are regularly formed nouns with the derivational suffix *-het* ('-ness'), e.g., *skvalleraktighet* (lit. 'gossipiness'), *skäggighet* (lit. 'beardiness'), *snubblighet* (lit. 'stumblyness') and *överklokhhet* (lit. 'over-wiseness'). Nevertheless, these words are extremely rare in modern Swedish texts.

Written Swedish is becoming more gender neutral, and consequently, many derivatives with the suffixes *-inna* and *-erska*, forming older feminine designations, have already been deleted from previous revisions of SAOL (see Gellerstam, 2009b). However, there are still a number of such derivatives among the non-matches, such as *givarinna* ('female donor') and *giftblanderska* ('female poison mixer').

The use of other lemmas in the list, such as *eskimåtröja* ('Eskimo sweater') and *fruntimmerssyssla* (lit. 'women's occupation'), has probably also declined because they are considered less appropriate in modern society, for various reasons. Another reason is that they may have acquired negative connotations (see Gellerstam, 2009b, pp. 82; Petersson & Sköldberg, 2020).

Finally, even though SAOL aims at reflecting the general Swedish language, a lot of specialized words occur in the lemma list, e.g., *dolicocephali* ('dolichocephaly') (in Figure 2). The fact that these words are non-matches in a corpus that even includes Wikipedia might indicate that those lemmas should not be a part of the SAOL lemma list either.

To sum up, it would appear that many of the non-matches could be excluded from the lemma list in SAOL without further ado, but some of the words among the non-matches may require further lexicographic investigation. The word *civiltjänstgörare* ('person doing civil service') is more common in Finnish-Swedish texts, and SAOL has also included a number of Finnish-Swedish words in recent editions (see Gellerstam, 2009b, pp. 76–77). A word like *bofinksbo* may have been included originally to illustrate that compounds with *bofink-* use *s* as a linking morpheme (see Berg et al., 2010, pp. 569–570).

It should also be noted that among the non-matches, there are Swedish words that are probably in use, but mainly in other contexts than those covered by the current corpora (e.g., *krassefrö* 'cress seed' and *vitlökskapsel* 'garlic capsule').

The use of these, as well as the other lemmas in the list of non-hits, will undergo further lexicographic investigations before a possible exclusion of obsolete words takes place for the forthcoming edition of SAOL.

If a lemma is deleted from an edition of the glossary, it will most certainly be searchable in the historical resource SAOLhist in the same way as previously excluded lemmas. A search in SAOLhist with, for example, the search string "%erska" shows several older feminine designations included in previous editions, among them *sanningssökerska* (lit. 'truth seekeress'), *kaffekokerska* (lit. 'coffee cookess') and *konstmakerska* (lit. 'art makeress'). Another search in the resource reveals that excluded SAOL lemmas can be included again. For example, the noun *aeronaut* ('aeronaut') and the adjective *aeronautisk* ('aeronautical') were deleted from the editions 11 (1986) and 12 (1998), but were included again in edition 13 (2006), probably due to their frequency in use.

5. Final Remarks

The revisional work for the 15th edition of SAOL is underway. The study presented here shows that approximately 3.5% of the headwords of the current edition, SAOL 14 (2015), do not occur in any inflectional form in our corpus of 3 billion words. The candidates for exclusions, therefore, amount to approximately 4,000 headwords.

Based on current corpora and modern language technology-based methods, the lexicographical decisions made by the SAOL lexicographers of today can be more efficient and built on a more solid scientific ground than was the case for previous editors.

SAOL needs to be systematically updated to maintain its role as a useful synchronic dictionary relevant for both linguists and non-linguists. For those users who are interested in the Swedish vocabulary of older times, other sources are available, e.g., SAOLhist.

The glossary only includes a selection of contemporary Swedish words, and this selection must be as representative as possible. It should also be noted that all versions of the same edition of SAOL (both print and digital) are expected to list the same headwords. Hence, lemma exclusion is an important building block for this purpose and a necessary complement to lemma inclusion.

Keeping the lemma list in SAOL up to date is also a prerequisite for new areas of use. SAOL could, among other things, strengthen its role as a useful resource in Swedish language technology.

References

- Berg, S., Holmer, L., & Sköldberg, E. (2010). Time to say goodbye? On the exclusion of solid compounds from the Swedish Academy Glossary (SAOL). In A. Dykstra, & T. Schoonheim (Eds.), *Proceedings of the 14th EURALEX International Congress* (pp. 567–576). Fryske Akademy.
- Berg, S., & Malmgren, S.-G. (2010). Varför travintresse ska ut och fotbollsintresse in. *Språkbruk* 2010(4). Retrieved July 30, 2024, from <https://sprakbruk.fi/artiklar/varfor-travintresse-ska-ut-och-fotbollsintresse-in/>
- Borin, L., Forsberg, M., & Roxendal, J. (2012). Korp – the corpus infrastructure of Språkbanken. In *Proceedings of LREC 2012* (pp. 474–478). ELRA.
- Diamond, G. (2016). Making Decisions about Inclusion and Exclusion. In P. Durkin (Ed.), *The Oxford Handbook of Lexicography* (pp. 532–545). Oxford University Press.
- Gellerstam, M. (2005). Ord som går i pension. Om arbetet med att rensa bort gamla ord i SAOL. In R. Vatvedt Fjeld, & D. Worren (Eds.), *Nordiske studiar i leksikografi* 7 (pp. 133–139). Nordisk foreining for leksikografi.

- Gellerstam, M. (2009a). Ordförrådets förändring i SAOL . In M. Gellerstam (Ed.), *SAOL och tidens flykt. Några nedslag i ordlistans historia* (pp. 85–104). Norstedts.
- Gellerstam, M. (2009b). SAOL i många upplagor. In M. Gellerstam (Ed.), *SAOL och tidens flykt. Några nedslag i ordlistans historia* (pp. 53–83). Norstedts.
- ten Hacken, P. (2020). Norms, new words, and empirical reality. *International Journal of Lexicography*, 33(2), 135–149.
- Hammarstedt, M., Schumacher, A. Borin, L., & Forsberg, M. (2022). *Sparv 5 User Manual. Research Reports from the Department of Swedish, Multilingualism, Language Technology*. Göteborg: University of Gothenburg 2022/02. Retrieved July 30, 2024, from <https://gupea.ub.gu.se/bitstream/handle/2077/73604/GU-ISS-2022-02%20Sparv%205%20User%20manual.pdf>
- Hanks, P. (2019). Lexicography. In R. Mitkov (Ed.), *The Oxford Handbook of Computational Linguistics*, 2 ed. (pp. 473–493). Oxford University Press.
- Holmer, L., Hult, A.-K., & Sköldberg, E. (2015). Spell-checking on the fly? On the use of a Swedish dictionary app. In I. Kosem et al. (Eds.), *Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of the eLex 2015 conference* (pp. 356–371). Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd.
- Holmer, L., Lillieström, A., Sköldberg, E., & Uppström, J. (2024). SAOL och svensk språkvetenskaplig infrastruktur – nu och i framtiden. In E. Volodina et al. (Eds.), *Proceedings of the Huminfra Conference (HiC 2024)* (pp. 68–75). LiU Electronic Press.
- Klosa, A., & Lungen, H. (2018). New German Words: Detection and Description. In J. Čibej, V. Gorjanc, I. Kosem, & S. Krek (Eds.), *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts* (pp. 559–569). Ljubljana University Press, Faculty of Arts.
- Petersson, S., & Sköldberg, E. (2020). To discriminate between discrimination and inclusion: A lexicographer's dilemma. In Z. Gavriilidou, M. Mitsiaki, & A. Fliatouras (Eds.), *Proceedings of the 19th EURALEX International Congress, Vol. 1*. (pp. 381–386). Democritus University of Thrace.
- Rundell, M. (2015). From Print to Digital: Implications for Dictionary Policy and Lexicographic Conventions. *Lexikos*, 25, 301–322.
- Sánchez Ibáñez, M., & Maroto, N. (2021). Beyond timelines: the challenges of combining theoretical premises and speakers' insights about the assessment, validation and inclusion of Spanish neologisms in dictionaries. *International Journal of Lexicography*, 34(3), 358–381.
- SAOL 11: *Svenska Akademiens ordlista*. (1986). 11th edition. Norstedts.
- SAOL 12: *Svenska Akademiens ordlista*. (1998). 12th edition. Norstedts.
- SAOL 13: *Svenska Akademiens ordlista*. (2006). 13th edition. Norstedts.
- SAOL 14: *Svenska Akademiens ordlista* (2015). 14th edition. Norstedts. Retrieved May 7, 2024, from <https://svenska.se/saol/>

SAOLhist: Välkommen till SAOLhist. Retrieved May 7, 2024, from <https://spraakbanken.gu.se/saolhist/>

Sköldberg, E., Holmer, L., Volodina, E., & Ildikó Pilán (2019). State-of-the-art on monolingual lexicography for Sweden. *Slovenščina 2.0*, 7(1), 13–24.

Svensson, A. (2019). Här är orden som försvinner. *Språktidningen* 2019(8). Retrieved May 7, 2024, from <https://spraktidningen.se/artiklar/har-ar-orden-som-forsvinner/>

Trap-Jensen, L. (2013). Researching lexicographical practice. In H. Jackson (Ed.), *Bloomsbury Companion to Lexicography* (pp. 19–30). Bloomsbury.

Vikør, L. S. (2009). Lexicography and language planning in Scandinavia and the Netherlands. In S. Nielsen, & S. Tarp (Eds.), *Lexicography in the 21st Century. In honour of Henning Bergenholtz* (pp. 123–143). John Benjamins.

Acknowledgements

This work has been funded by the Swedish Academy, and by Nationella språkbanken – jointly funded by its 10 partner institutions and the Swedish Research Council (2018–2024; dnr 2017-00626).

Contact information

Louise Holmer

University of Gothenburg, Department of Swedish, Multilingualism, Language Technology
louise.holmer@svenska.gu.se

Ann Lillieström

University of Gothenburg, Department of Swedish, Multilingualism, Language Technology
ann.lilliestrom@gu.se

Emma Sköldberg

University of Gothenburg, Department of Swedish, Multilingualism, Language Technology
emma.skoldberg@svenska.gu.se

Jonatan Uppström

University of Gothenburg, Department of Swedish, Multilingualism, Language Technology
jonatan.uppstrom@svenska.gu.se

Kathryn Hudson

THE BENEFITS OF BIO(LEXICOGRAPHY)

A Topical Approach to Lexicographic Practice

Abstract This paper combines insights drawn from the author's ongoing linguistic and ethnographic work in the Americas and Micronesia with lessons from Lichtenberk's (2003) work on creating dictionaries for languages in transition to explore the utility of biolexicography and other topical lexicographic approaches. Particular attention is given to key features of biolexicography – including biolexica, the role of communities of practice, the use of ethnographic methodologies, considerations of identity, and linguistically-mediated ecological engagement – and to the relevance of this and other topical lexicographic approaches for languages in transition and their speakers. The ways in which these approaches are shaped by the lexicographic inheritance and by engagements with languages experiencing attrition or change are also considered. This discussion illustrates that the utility of biolexicography is rooted in the linguistic and sociocultural significances of biolexica in ways that integrate lexicographic products into broader cultural and linguistic systems and center speaker communities in the lexicographic process.

Keywords ecology; ethnography; language change; lexicography

1. Introductory Remarks

Topical or thematic lexicographic materials focus, exclusively or primarily, on a particular subject matter recognized in a particular sociocultural context or on a cluster of topics closely related within a particular cultural tradition. They are often conceptualized as extensions of – or supplements to – more generalized lexicographic productions rather than as a potentially primary methodology; however, their possibly utility is far greater than these classifications imply. This is particularly true for languages experiencing attrition or change and for those facing assimilationist pressures from colonial linguistic systems. It is also true for languages spoken by communities in the process of (re)asserting or (re)establishing their identities, particularly in postcolonial situations.

This paper develops a conceptual and practical examination of biolexicography – an ethnographically-rooted and ecologically-based topical methodology for community driven, use-oriented lexicographic work – to explore this and other forms of topical lexicography. It combines lessons from Lichtenberk's engagement with the complexities of creating dictionaries for languages in transition with insights drawn from the author's ongoing linguistic and ethnographic work in the Americas and Micronesia to explore the practicality and utility of topical lexicographic approaches. Particular attention is given to (i) the relevance of this approach for languages in transition and their speakers, and (ii) the necessity of an ethnographic foundation that is critically oriented and collaboratively developed.

2. The Lexicographic Inheritance and Language Change

Throughout much of its long history, lexicography has held “a mediating position for language use” focused on “[e]xplaining foreign languages or areas of native languages which remain foreign...[or] unintelligible to the common speaker and writer” (Hüllen, 2000, p. 6). This is, in part, a product of its developmental history and trajectory. It is also reflective of the close association of lexicography with academic linguistics and of the authority that is frequently ascribed to – or claimed by – academic researchers. By explaining language with and through language (see Hüllen, 2020) and using lexicographic presentations to facilitate “correct and successful [linguistic] performance” (Hüllen, 2020, pp. 4–5), lexicography eventually emerged as a subdiscipline of linguistics that shares “basic assumptions about language with other subdisciplines” (ibid.) and is central to many documentation efforts. The lexicographic inheritance and issues of language change helped motivate this process; considering their role in the development of lexicography is thus crucial to any consideration of the development and implementation of more specific lexicographic approaches.

2.1 The Lexicographic Inheritance

Biocentricity – like other topical lexicographic approaches – represents a critical and highly contextualized response to the lexicographic inheritance. The earliest known lexicographic materials were developed in ancient Mesopotamia, where monolingual Sumerian wordlists were used for training scribes by c. 3200 BCE and bilingual Sumerian-Eblaite wordlists existed by c. 2400 BCE (Considine, 2015, p. 605). Similar materials appeared in Egypt by the 18th century BCE, in India around 300 BCE, in Greece during the 3rd century BCE, and in Chinese compendium of glosses known as the *Erya* (ibid). Dictionaries of a relatively modern form first appear in China with Yang Xiong’s dialect dictionary c. 18 AD and Xu Shen’s dictionary of Chinese characters compiled c. 149 AD (Considine, 2015, pp. 605–606). The Western tradition arguably originated with Hesychius of Alexandria’s Greek dictionary in the 5th – 6th century AD; other early lexicographic traditions appeared in the Indian subcontinent in the 6th century AD, in the Arab world c. 791 AD, in Tibet c. 814 AD, and in the Hebrew-speaking world c. 902 (ibid).

The strongly academic orientation of these early materials continued into the modern era. Early editions of the Oxford English Dictionary, for example, were developed through “impersonal processes” that produced “a more strictly descriptive format” (Kistner, 2013, p. 802). This reflected a tradition of defining dictionaries as catalogues of words and definitions. It was also a product of the persistent belief that dictionaries can provide “accurate and complete accounts of a community’s (standard) language... comprised of words which have specific meanings” (Seargeant, 2011, p.1) – a view that conceptually linked dictionaries with comprehensiveness and with assumed academic (i.e., etc) abilities to identify, define, and record this scope.

This perspective is predicated on a trio of interrelated beliefs: (i) words form the core units of language, (ii) all words possess fixed and delineable meanings, and

(iii) a language is a defined system composed of a vast but defined set of words and their meanings (Sergeant, 2011). Dictionaries are thus tasked with recording languages “in the *fullest possible* detail” (Sergeant, 2011, p. 3, emphasis added) and viewed as “authorit[ies] that tell [] us whether a certain locution is actually part of the language” (Dolezal, 2006, pp. 695). The processes of their creation are similarly imbued with legitimizing power, since dictionaries and dictionary-making often symbolically function as evidence of linguistic realness and validity among both speakers and outside observers (see e.g., Corris et al., 2000). Lexicographers thus have a tendency to “want to include as much information as is known” (Corris et al. 2000, 341), and lexicographic work – particularly in contexts of endangerment and attrition – often seeks to preserve a language in its entirety in dictionary form “for future study of revival...[by] linguists and researchers” (Corris et al., 2000, pp. 341, 330). However, “this wealth of information can be counterproductive for users” (Corris et al., 2000, p. 341).

Despite these consequences, the dominance and relative inflexibility of these beliefs has also encouraged the development of new and more customized approaches to dictionary creation. This is particularly true in field-based work and in cases of language attrition and change, since these situations do not always – or often – fit neatly within traditional lexicographic methods and formats (see e.g., Cablitz, 2011; Haviland, 2006; Mosel, 2004, 2011; Vamarasi, 2014). It is also apparent in more general discussions about the formatting of dictionary entries (see e.g., Atkins & Rundell, 2008; Field, 2009; Ivanishcheva, 2016; Kroskrity, 2015), the treatment of loanwords (see e.g., Crowley, 1992; Kroskrity, 1993, 1998; Lichtenberk, 2003), and the particularities of orthographic representation (see e.g., Mosel, 2004). Considered together, these discussions reflect a growing awareness of the need for contextualized approaches sensitive to the needs and experiences of individual communities.

2.2 Lexicography and Language Change

Consideration of – and engagement with – languages undergoing attrition or change further illustrates the need for contextualized lexicographic approaches sensitive to the needs and experiences of individual communities of linguistics practice. In these cases, community conceptualizations of what constitutes fluency – and of what makes someone a speaker – often differ markedly from those of academic linguists and lexicographic researchers. Knowledge of indigenous words is often viewed as evidence of speaker status, even if those words are not actively used or are used only in the context of discourse that primarily utilized a different linguistic tradition. The identities of these words reflect a community’s culture and their sociocultural, economic, and political positions within a broader postcolonial landscape. They are also a product of cultural and linguistic politics and values, in academia and beyond, and of agentive processes of identity (re)assertion.

The complex and contextualized nature of language attrition and change make it necessary to critically engage with the lexicographic process. This necessarily begins with audience identification and with the prioritization of different audiences. There

are two general audiences relevant to any lexicographic production: (1) an emic audience made up of speakers and/or members of descendent communities, and (2) an etic audience consisting of academics and other outsiders with an interest in a particular linguistic system. The emic audience should, through linguistic patrimony and ownership, be given priority, though etic audiences are often given equal or greater consideration. Lichtenberk (2003) devoted considerable attention to how his dictionary of Toqabaqita would be used. He identified two groups of users – local Toqabaqita people and academics – and believed his dictionary could address the needs and desires of both audiences. However, his concern with the limitations of emic audiences and his belief that it is “unrealistic to think that the dictionary will be frequently used by the local people” because its main value for them will be the recognition garnered by its existence suggest an implicit – and cautionary – focus on etic users (Lichtenberk, 2003, pp. 398–400).

Processes of lexicographic selection and orthographic standardization are equally crucial. Word selection should be culturally motivated, contextualized, and reflective of the language as it exists within a particular speaker community. Loanwords regularly used by speakers should be included regardless of their origins; those that co-occur with indigenous alternatives should be recorded as distinct synonyms. In all cases, the orthography used for these words should match conventions used by speakers, even when they appear illogical or imprecise to professional observers, and a user-friendly orthographic system should be developed in collaboration with community members when no pre-existing system exists. Integrating these practices increases the contextual accuracy and appropriateness of lexicographic materials while also making them more reflective of emic speaker perspectives (see e.g., Haviland, 2006; Pawley, 2001). Selectively applying them, however, can limit the accessibility of lexicographic materials – particularly for emic users – and exacerbate issues in the lexicographic process. These include notions of ‘pure’ or ‘good’ linguistic forms, resistance to lexical and phonological change, a desire to edit words deemed overly reflective of a language’s colonial experiences, and qualitative judgments of the language used by speakers.

In Lichtenberk’s (2003) work with Toqabaqita, for example, issues of lexical variability – particularly loanwords – and orthographic representation are central foci. His treatment of loanwords is based on the degree to which loanwords fit Toqabaqita phonological and phonotactic patterns rather than on the use patterns of these words by speakers (*ibid.*). Loanwords are included if they fit these patterns or have been modified to fit them, though words without modification are still excluded if they belong to a set in which other words are not modified (e.g., numbers or day names) or if they only occurred once (Lichtenberk, 2003, pp. 394–397). Lexical variability is thus governed by phonological and phonotactic parameters rather than by patterns of use, motivating an academically oriented definition of the Toqabaqita lexicon. Lichtenberk’s (2003) orthographic conventions improve the consistency and phonetic accuracy of the writing system but differ – sometimes significantly – from those used by Toqabaqita people, thus further reducing their ability to utilize the dictionary.

3. Biolexicography: A Multispecies, Use-Driven Approach

Biolexicography is an ethnographically rooted and ecologically based topical methodology for contextual and use-oriented lexicographic work being developed through ongoing collaborative and community-driven work with communities in Latin America and western Micronesia.¹ It represents a highly contextualized response to the orthodox beliefs imbedded in the lexicographic inheritance and to the pressures of language attrition and change in postcolonial communities. It also reflects the reconceptualization of culture, identity, and language within these communities and acknowledges the complex relationships that exist between these categories and the more-than-human world. The utility of biolexicography is rooted in the linguistic and sociocultural significances of words pertaining to animal and plant species – collectively described as biolexica – and in its emphasis on communities of practice, its ethnographic foundations, and its cognizance of identity (re)formulation in postcolonial situations. Each of these factors will be considered in turn.

3.1 Biolexica

Within the framework of biolexicography, words pertaining to animal and plant species are collectively known as biolexica. The use of this term instead of more semantically restrictive labels such as zoonyms allows the inclusion of words identifying both plant and animal species, thus allowing the methodology to better align with systems of traditional ecological knowledge and classification. It also facilitates the incorporation of words describing (i) parts of animals and plants, (ii) aspects of the life cycles and behaviors of animals and plant species, (iii) species or species groupings, and (iv) human behaviors related to or based on other species. Biolexica thus catalogue patterns of human interaction with more-than-human worlds – including, among other things, culturally defined relationships between animals and space, conceptualizations of wild versus domestic, cultural typologies of the natural world, and the potential relationships between humans and animals – and record traditional ecological knowledge. They also aid in the reconstruction of precolonial environments, foster ecological engagement in contemporary communities, and can be used to reassert or reestablish real or symbolic connections to traditional homelands and environments. By extension, biolexica can also motivate or maintain interest in protecting and maintaining these spaces and their diverse inhabitants.

3.2 Communities of Practice

Lexicographic work should reflect a specific and well-defined context since dictionaries are “artifact[s] *designed* with care to fit *precise specifications*” (Atkins, 2008, p. 35). These specifications are determined by the relevant community of practice – defined here as a speaker community or descent community with interests in developing, maintaining, and/or using a particular language – and guided by their desires, needs, and interests. Such community-driven work requires awareness of, and sensitivity to,

¹ Community elders and other community members request that their identities and specific locations remain anonymous as a condition of access and ongoing work.

the contexts of these communities – cultural, economic, historical, political, and social – and the ways in which they view and use their language within them. This allows the parameters of linguistic use to be defined by speaker communities in contextually appropriate ways that encourage community engagement with lexicographic work. It also facilitates the integration of dictionaries and other lexicographic products into broader cultural and linguistic systems in ways that increase their relevance and utility for speakers.

Such purpose-driven lexicography is particularly significant for speakers of languages undergoing attrition or change. Although many lexicographers working with these languages agree that “dictionaries for endangered languages...differ mostly in the amount of information in certain parts of the microstructure,” (Corris et al., 2000, p. 2), there are others (e.g., Cablitz, 2011; Field, 2009; Ivanishcheva, 2016) who advocate for more contextually and culturally sensitive approaches in which these materials also differ in the nature of their content and structure. Biolexicography is one such approach; its focus on biolexica significant in a particular community of practice leads to distinct lexicographic content structured around culturally significance relationships between people, language, and environment. By reflecting the cultural, linguistic, and sociopolitical worlds of a community, biolexicography and other contextualized approaches capture the “cultural constructs embedded and reflected in the...language” (Field, 2009, p. 296). They also preserve “unique facts about the culture, and language as part of the culture, of an indigenous minority” (Ivanishcheva, 2016, p. 83) and incorporate the perspectives of the speaker community.

3.3 Ethnographic Foundations

Biolexicography, like many topical lexicographic approaches, is strongly rooted in ethnographic fieldwork since “a great deal of cultural knowledge is part of the linguistic competence of speakers” (Cablitz, 2011, p. 447). Its methods are culturally motivated and contextualized; its products incorporate cultural knowledge and practices in ways that articulate the sociocultural contexts of a language with the historical, economic, and sociopolitical landscapes of its speakers. The integration of lexicographic and ethnographic methodologies is thus crucial. Simply being in the field for purposes of data acquisition does not allow for the development of nuanced and contextualized understandings of words, their meanings, or their uses; engagement with, and participation in, the field are required. This is particularly true of biolexicography, since identifying and understanding biolexica – and, crucially, their relationships to other cultural and linguistic systems – necessitates engagement with contexts in which the traditional ecological knowledge associated with these words is made manifest through the intersection of sociocultural and linguistic behaviors. When ethnographic fieldwork is foundational to lexicographic endeavours in this way, the facts that “[c]ulture permeates lexicons” and “cultural traditions are sometimes inseparable from linguistic form” are recognized in ways that benefit languages and their speakers (Field, 2009, p. 300). Ethnographic perspectives thus facilitate the lexicographic selection of words and help define the foci of topical

lexicographic approaches: they reveal which words are active participants in speakers' cultural lives, which relate to constructions and conceptualizations of identity, and – by extension – which are *functioning* parts of the language as it is used in a particular community.

3.4 Implications for Identity and the More-Than-Human World

Biolexica – like the lexicographic materials that record them – provide a means of asserting cultural and linguistic competency and identity through affiliation with particular environments, even when more expansive knowledge of cultural and linguistic systems has been lost or eroded. The knowledge and use of biolexica reflect an individual's access to the cultural knowledge associated with them and, by extension, indicate sufficient engagement with the relevant ecological systems to use these words appropriately. They are also often viewed as evidence of a connection with history and with the legitimacy that derives from it. Postcolonial indigenous communities often feel “a...need to demonstrate that they exist, have been present for generations, and have co-shaped the culture and cultural landscape of a certain place” (Jordan, 2012, p. 129); this need can be addressed through use of biolexica recorded in lexicographic materials and through access to – or possession of – these materials. By linguistically linking themselves with their environment in this way, communities more permanently connect themselves to particular locations and provide evidence of historical existence.

Biolexicography also has implications for conservation efforts and sustainability initiatives, which frequently parallel efforts for maintain and preserve cultural identities. Biodiversity is generally more stable in areas managed by indigenous populations, and the practices that facilitate this – including the creation of diverse and species-rich spaces, the active management of land, the pursuit of ecological restoration, and the diachronic monitoring of ecosystems – are often rooted in lexicographically preserved systems of traditional ecological knowledge. The ethnographic elicitation and lexicographic recording of biolexica can trigger these efforts, guide their methodologies, and contribute to their maintenance, particularly when cultural and practical information is recorded in the definitions, etymologies, example sentences, and/or supplemental discussions that accompany them.

4. Concluding Remarks

Biolexicography provides an ethnographically rooted and ecologically-based topical methodology for community driven, use-oriented lexicographic work that can be used to explore the utility of topical lexicography. The lexicographic inheritance and engagement with issues of language attrition and change motivated its development; their role in the development of lexicography more generally provides a useful framing for considerations of this and other topical lexicographic approaches. The utility of biolexicography is rooted in the linguistic and sociocultural significances of biolexica; its emphasis on communities of practice, ethnographic foundations, and connection to processes of identity (re)formulation facilitates the integration of its

products into broader cultural and linguistic systems. The ongoing development of this and other topical approaches to lexicography will continue to illustrate their significance, particularly in contexts of attrition and change, while also centering speaker communities in the lexicographic process.

References

- Atkins, S. (2008). Theoretical Lexicography and its Relation to Dictionary-making. In T. Fontenelle (Ed.), *Practical lexicography: A Reader* (pp. 31–50). Oxford University Press.
- Cablitz, G. H. (2011). Documenting Cultural Knowledge in Dictionaries of Endangered Languages. *International Journal of Lexicography* 24(4), 446–462.
- Considine, J. (2015). A Chronology of Major Events in the History of Lexicography. In P. Durkin (Ed.), *The Oxford Handbook of Lexicography* (pp. 605–624). Oxford University Press.
- Corris, M., Manning, C., Poetsch, S., & Simpson, J. (2000). Dictionaries and Endangered Languages. In D. Bradley, & M. Bradley (Eds.), *Language Endangerment and Language Maintenance: An Active Approach* (pp. 329–348). Taylor & Francis Group.
- Crowley, T. (1992). *A Dictionary of Paamese*. Australian National University.
- Dolezal, F. (2006). World Englishes and lexicography. In B. Kachru, Y. Kachru, & C. Nelson (Eds.), *The Handbook of World Englishes* (pp. 694–708). Blackwell.
- Duval, T., & K. Kuiper. (2001). Maori Dictionaries and Maori Loanwords. *International Journal of Lexicography* 14, 243–60.
- Field, M. C. (2009). Metaphor, Mythology, and the Navajo Verb: The Role of Cultural Constructs in the Lexicography of Endangered Languages. *Anthropological Linguistics* 51(3/4), 296–302.
- Haviland, J. (2006). Documenting Lexical Knowledge. In J. Gippert, N. P. Himmelmann, & U. Mosel (Eds.), *Essentials of Language Documentation* (pp. 1–30). Mouton de Gruyter.
- Hüllen, W. (2000). *English Dictionaries, 800-1700: The Topical Tradition*. 1st ed. Oxford University Press, Incorporated.
- Ivanishcheva, O. N. (2016). Dictionaries of Critically Endangered Languages: Focus on Users. *Journal of Linguistics* 67(1), 73–86.
- Kistner, K. (2013). ‘A Word Factory was wanted’: Organizational objectivity in the making of the Oxford English Dictionary. *Social Studies of Science* 43(6), 801–828.
- Kroskrity, P. V. (1993). *Language, history, and identity: Ethnolinguistic studies of the Arizona Tewa*. University of Arizona Press.
- Kroskrity, P. V. (2010). Language ideologies – evolving perspectives. In Ju. Jaspers, J.-O. Östman, & J. Verschueren (Eds.), *Society and language use, Handbook of Pragmatics Highlights, Vol. 7* (pp. 192–211). John Benjamins.

Kroskrity, P. V. (2015). Designing a Dictionary for an Endangered Language Community: Lexicographical Deliberations, Language Ideological Clarifications. *Language Documentation & Conservation*, 9, 140–157.

Lichtenberk, F. (2003). To List or Not to List: Writing a Dictionary of a Language Undergoing Rapid and Extensive Lexical Changes. *International Journal of Lexicography* 16(4), 387–401.

Mosel, U. (2004). Dictionary Making in Endangered Speech Communities. In P. K. Austin (Ed.), *Language Documentation and Description*, Vol. 2 (pp. 39–54). School of Oriental and African Studies.

Mosel, U. (2011). Lexicography in Endangered Language Communities. In P. K. Austin, & J. Sallaback (Eds.), *The Cambridge Handbook of Endangered Languages* (pp. 337–353). Cambridge University Press.

Pawley, A. (2001). Some Problems of Describing Linguistic and Ecological Knowledge. In L. Maffi (Ed.), *On Biocultural Diversity: Linking Language, Knowledge, and the Environment* (pp. 228–247). Smithsonian Institution Press.

Seargeant, P. (2011). Lexicography as a philosophy of language. *Language Sciences* 33, 1–10.

Vamarasi, M. (2014). The creation of learner-centered dictionaries for endangered languages: a Rotuman example. *Journal of Multilingual and Multicultural Development* 35(2), 167–178.

Contact information

Kathryn Hudson
University at Buffalo
khudson@buffalo.edu

Boris Kern

CONSIDERING WORD FORMATION IN COMPILING DICTIONARIES

Abstract In this article, we aim to illustrate, using the example of Slovenian, which aspects need to be considered when compiling a new dictionary. In compiling dictionaries, considering word formation is crucial in several phases: a) selecting criteria for including words in the dictionary, b) uniformity in dictionary definitions of headwords that have some word-formation relation, and c) consistency in qualifying certain senses as stylistically marked. Regarding the criteria for including words in the dictionary, this article pays special attention to statistical analysis of the correlation between multistage derivatives and the frequency of individual words in word-formation chains, in corpus sources, and in the case at hand the Gigafida 2.0 and metaFida corpora. It is assumed that the frequency of a word decreases as its formation stage becomes higher. This article presents dictionary solutions as indicated by dictionary concepts for the monolingual dictionary of standard Slovenian.

Keywords word formation; multistage word formation; Slovenian lexicography

1. Introduction

Like other Slavic languages, Slovenian lexemes are characterized by an extremely rich morphemic structure, which is a result of multistage word formation. For example, in the first stage, the adjective *mlad* ‘young’ yields the noun *mladost* ‘youth’, which in turn yields the adjective *mladosten* ‘youthful’ in the second stage, which in turn yields the noun *mladostnik* ‘adolescent’ in the third stage, which in turn yields the possessive adjective *mladostnikov* ‘adolescent’s’ in the fourth stage (Skarżyński, 1999; 2003; Olejniczak, 2003; Kern, 2010; 2017; Vidovič Muha, 2018).¹ This linguistic fact influences certain aspects of compiling dictionaries for Slavic languages at several stages:² a) selecting criteria for including words in the dictionary, b) uniformity in dictionary definitions of headwords that have some word-formation relation, and c) consistency in qualifying certain senses as stylistically marked (Skarżyński, 1999; 2003; Olejniczak, 2003; Burkacka, 2012; Kern, 2010; 2017). Regarding the criteria for including words in the dictionary, it is necessary to define strategies for cases

¹ From 2021 to 2024, the project Formant Combinatorics in Slovenian is being carried out in collaboration with ZRC SAZU, the Jožef Stefan Institute, and the University of Maribor. The objective of the project is to explore the combinatorics of word-formation formants, which will make it possible to present the characteristics of word formation and semantic extension mechanisms of Slovenian on contemporary language material. The language technology objective of the project is a pioneering creation of the first training set and the first language technology application allowing automatic morpheme segmentation of Slovenian words (see <https://isjfr.zrc-sazu.si/en/programi-in-projekti/formant-combinatorics-in-slovenian>).

² It is worth noting that from the sixteenth to eighteenth century—the period when efforts to compile monolingual dictionaries emerged—the “word-family” arrangements of lexemes in dictionaries was a serious competitor to the alphabetical arrangement of lexemes (August, 1993, pp. 188–189).

in which so-called word-formation gaps or hypothetical derivations³ occur in the word-formation chains. A particular element of a chain may not be present in the language. For example, in the Slovenian word-formation chain *okus* 'taste' → *okusiti* 'to taste' *pf.* → *okuševati* 'to taste' *impf.* → *okuševalec* 'taster', the verb form *okuševati* occurs only once in the Slovenian reference corpus Gigafida 2.0. For this reason, it is not included in dictionaries of Slovenian. Hypothetical derivatives can also be observed at the end of word-formation chains, where, according to previous studies, diminutives, possessive adjectives, adjectives ending in *-ni* or *-ski/-ški*, nouns ending in *-ost*, and verbal nouns (e.g., *zaspanost* 'sleepiness') occur most frequently in Slovenian. Regarding qualifying the senses as stylistically marked, it should be added that it is of course neither necessary nor usual for all words in a word-formation chain to have the same connotative meaning. However, it is useful in lexicographic work to review all multistage derivatives, both in terms of qualification and dictionary definitions.

In dictionary definitions, it is sensible to use unified classificatory (*genus proximum*) and distinctive semantic components (*differentia specifica*) within word-formation chains where possible (e.g., **vegán** kdor ne je hrane živalskega izvora in zavrača vsakršno izkoriščanje živali → **vegánka** ženska, ki ne je hrane živalskega izvora in zavrača vsakršno izkoriščanje živali → **vegánstvo** prehranjevanje s hrano, ki ni živalskega izvora, in zavračanje vsakršnega izkoriščanja živali; **vegan** someone who does not consume animal-based food and rejects any form of animal exploitation → **female vegan** a woman who does not consume animal-based food and rejects any form of animal exploitation → **veganism**: eating food that is not of animal origin and rejecting any form of animal exploitation').⁴ It is also important to maintain a unified semantic structure, in terms of the relationship between the primary meaning and subordinate meaning. An example of inconsistency can be seen with the verbs *videti* 'to see,' *slišati* 'to hear,' and *poslušati* 'to listen' in the Dictionary of the Slovene Standard Language, 2nd edition (2024): in the second meaning of *videti* 'to see' and *slišati* 'to hear,' both express 'becoming aware through sight or hearing,' whereas the verb *poslušati* 'to listen' has the same semantic transition into the field of mental perception expressed in a subordinate meaning ('to perceive and understand the content of something by listening'). These three examples are systematically similar, so the semantic differentiation should be unified, especially in cases involving derived terms.

A methodology that seems to be particularly useful for analyzing such material, in all three phases listed above, is an analysis based on multistage word formation. This methodology extends the binary relation between the motivating and motivated word to larger groups of words, which include both direct derivatives (e.g., the noun *mladost* 'youth' with regard to the adjective *mlad* 'young') and indirect derivatives (e.g., the possessive adjective *mladostnikov* 'adolescent's' with regard to the noun *mladost* 'youth').⁵

³ The hypothetical derivative is a derivative whose introduction within word-formation chain is necessary especially in order to explain a higher-stage derivative that is attested in use.

⁴ Examples are from the eSSKJ: Dictionary of the Slovene Standard Language (2016–).

⁵ In this context, the studies by Čibej (2021) and Dobrovoljc et al. (2017) are also relevant for Slovenian.

2. Word-Formation Aspects in Dictionary Concepts for Slovenian

In Slovenian linguistics, two dictionary concepts were published approximately simultaneously: the first was a proposal for a modern Slovenian dictionary (Krek et al., 2013; hereinafter: Concept I), followed by a concept for a new explanatory dictionary of standard Slovenian by lexicographers at the Lexicological Section of the ZRC SAZU Fran Ramovš Institute of the Slovenian Language (Gliha Komac et al., 2015; hereinafter: Concept II).

2.1 Concept I: A Modern Slovenian Dictionary

Concept I undoubtedly presents a well-thought-out approach to the dictionary in a manner that follows modern lexicographic trends. However, the concept pays relatively little attention to the presentation of word-formation data. It mentions “word formation–related forms,” noting that it is necessary to consider how to address the problem of word formation–related forms in relation to polysemy (e.g., in the case of diminutives; Krek et al., 2013, p. 73), or the problem of asymmetry between formation-indicated diminutiveness and semantic non-diminutiveness, such as *jopica* ‘small sweatshirt’ versus *marelica* ‘apricot’ (and not ‘small umbrella’; Krek et al., 2013, pp. 74–75). The concept also states that the dictionary lists words as independent headwords when they feature competing word-formation suffixes that have had different normative values in previous linguistic resources (*-lec* vs. *-vec*; *zajedavec* vs. *zajedalec* ‘parasite’). The principle accepted is that, if both forms exceed the frequency threshold for inclusion in the lexical database, the words are treated as independent headwords “since it is possible to anticipate both semantic differentiation and variation in textual context and usage, which is usually conditioned by the frequency of each form.” The concept also mentions that the lexical database does not treat prefixes such as *evro-*, *pro-*, *anti-*, *e-*, and so on as independent headwords,⁶ but either as standalone headwords (e.g., *evro* ‘euro’) or as whole words to which such prefixes are attached (e.g., *biomasa* ‘biomass’; Krek et al., 2013, p. 75). The dictionary was not realized over the following years; however, due to the concept’s innovativeness, it undoubtedly deserves to be included in this research.

2.2 Concept II: A New Explanatory Dictionary

On the other hand Concept II, which was approved and adopted in 2015 by the Research Council of the ZRC SAZU Fran Ramovš Institute of the Slovenian Language, the ZRC SAZU Research Council, the ZRC SAZU Class of Philological and Literary Sciences, and the SAZU Presidency (eSSKJ 2017, p. 1), is a basis for the third edition of eSSKJ *Slovar slovenskega knjižnega jezika* (Standard Slovenian Dictionary, hereinafter: eSSKJ),⁷ which has been available as a growing dictionary⁸ on the lexicographic portal Fran since 2016. It is a new informative–normative explanatory dictionary of standard

⁶ This was the practice in some other Slovenian dictionaries: *Slovar slovenskega knjižnega jezika* (Standard Slovenian Dictionary, 1970–1991, 2014) and *Slovenski pravopis* (Slovenian Normative Dictionary, 2001).

⁷ The first edition was published between 1970 and 1991, and the second in 2014.

⁸ New dictionary entries are added each year; the exception was 2020.

Slovenian, with dictionary entries created entirely anew based on the analysis of contemporary language material. eSSKJ is the central lexicographic project of the ZRC SAZU Fran Ramovš Institute of the Slovenian Language and is planned to be a dictionary with approximately 100,000 dictionary entries, introduced at the macrostructural level by single-word headwords (Gliha Komac et al., 2015, p. 3). So far, 3,432 entries have been published in the dictionary.

The eSSKJ dictionary includes a word-formation tab, which, according to Concept II, is shown only for noun entries. In this tab, possessive adjectives, diminutives, feminatives, and masculatives are listed with hyperlinks if they meet frequency criteria and thus appear as independent entries.⁹ The terms feminative and masculative are used in a broad sense, encompassing examples of parallel derivatives (e.g., *igralec* ‘actor’ vs. *igralka* ‘actress’),¹⁰ but the terms themselves are not listed in the tab; instead, they are replaced by the labels *ženska oseba* ‘female’ or *moška oseba* ‘male’ before the related entry (see Figure 1). Feminatives are listed for paired masculine nouns on Fran, but not vice versa.

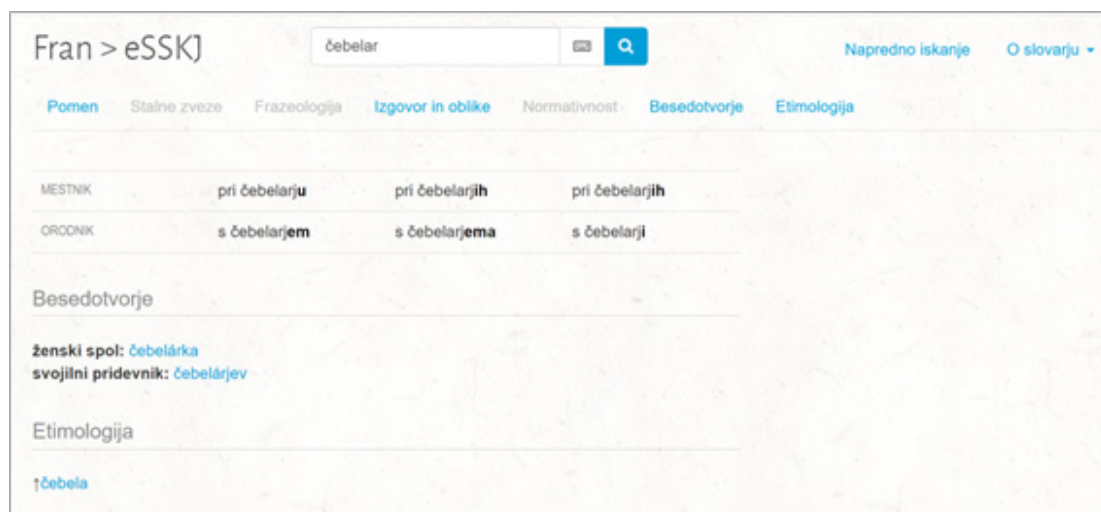


Fig. 1: Word formation tab for the entry *čebelar* ‘beekeeper’.

The word-formation tab also includes pairs such as *babica* ‘grandmother’ versus *dedek* ‘grandfather’. In these cases, the referencing is mutual; the female part of the

⁹ On the other hand, in the new Croatian dictionary *Hrvatski mrežni rječnik – Mrežnik*, the word-formation data are much more comprehensive. *Mrežnik* is a monolingual dictionary of standard Croatian, created from 2017 to 2021 as part of the research project of the same name, funded by the Croatian Science Foundation. Since 2022, the dictionary work has continued as an internal project of the Croatian Language Institute. The dictionary is available on the website www.rjecnik.hr/mreznik, and it includes three modules: a) a module for adult Croatian speakers, b) a module for elementary school students, and c) a module for speakers of Croatian as a second and foreign language. The module for adult Croatian speakers includes a section on word formation, which features the segmentation of a word into word-formation morphemes (e.g., *fizik-jar* for the entry *fizičar* ‘physicist’; phonetic changes are not indicated) and derivatives from the given entry (e.g., *fizičarev* ‘physicist’s’, *fizičarski* ‘physicist’s’, *fizičarka* ‘female physicist’). This is not limited to noun derivatives, as is the practice in the eSSKJ dictionary. Certain derivatives are also listed in the normative note tab, specifically when derivatives with different morphemes have different normative statuses (e.g., *čitalac* and *čitatelj* for ‘reader’). More about the principles of listing word-formation data can be found in the volume *Hrvatski mrežni rječnik – Mrežnik* or at <http://ihjj.hr/mreznik/uploads/5fcd02949bd03f2c37444dce7de2bfc.pdf> (Hudeček et al., 2016).

¹⁰ Both are direct derivatives of the verb *igrati* ‘to play’.

pair is listed in the word-formation tab of the masculine noun and vice versa. The same mutual connection exists for pairs for which the female derivative existed first due to social circumstances (e.g., *babica* ‘midwife’ vs. *babičar* ‘accoucheur’, *hostesa* ‘hostess’ vs. *hostesnik* ‘host’).

The lexicographer’s team decided to lower the frequency criteria for the inclusion of feminatives in the dictionary due to social responsibility. Feminatives (and masculatives) appear as independent entries when there is at least one occurrence in the corpus that can be used to create a dictionary example. For inclusion in the word-formation tab, one occurrence in the corpus is sufficient. The same frequency criteria apply for listing possessive adjectives and diminutives as independent entries as for other words. They are included in the word formation tab if they have at least twenty-five occurrences. Certain deviations are made for systemic alignment in cases of equivalent duplicates.

According to Concept II (Gliha Komac et al., 2015, p. 70), other primary derivatives from entries of different word classes can also be included in the dictionary if the analysis of the material shows that there are justified reasons for this (e.g., users’ communication difficulties, normative aspects, etc.). However, this practice was not established in the dictionary’s compilation. It is worth noting that the graded forms of adjectives, adverbs, and non-finite verb forms are systematically treated within inflectional-accentual patterns because they pertain to inflection rather than word formation. Data on secondary derivatives related to the entry are provided in the dictionary entries, where primary derivatives appear as headwords. Thus, the word formation tab shows at most one additional word-formation level. Exceptions are derivatives for which the intermediate formation stage exists only potentially but is not realized in the text; for example, *opičnjak* ‘ape-man’ will be listed under *opica* ‘ape’, not under the potential adjective **opični*.

Concept II also states that word-formation data, specifically the normative value of individual forms, are included in the normativeness tab (Gliha Komac et al., 2015, pp. 5–6). Examples include adjectival uses of nouns such as *aikido* (e.g., *aikido klub* ‘aikido club’), *kvir* (e.g., *kvir teorija* ‘queer theory’), and *reiki* (e.g., *reiki terapevt* ‘reiki therapist’), which are very common in contemporary material but are still understood by traditional normative studies as compounds rather than phrases. In the normativeness tab, there is a note: “The material shows that words like *video* are also used as adjectives (*video posnetek* ‘video recording’). Such examples are recognized in the 2001 Slovenian normative guide (§§ 487, 495, 498) as the first parts of compounds, and so writing them together (e.g., *videozapis* ‘video note’, *videoblog* ‘video blog’) is recommended, although separate writing is also allowed as non-preferred.” (More on this issue can be found in Gložančev, 2012, and Kern, 2012.)

In the future, word-formation data should definitely be included in the eSSKJ dictionary for other parts of speech, not just for nouns (similar to the way it is addressed in the Croatian Mrežnik dictionary).

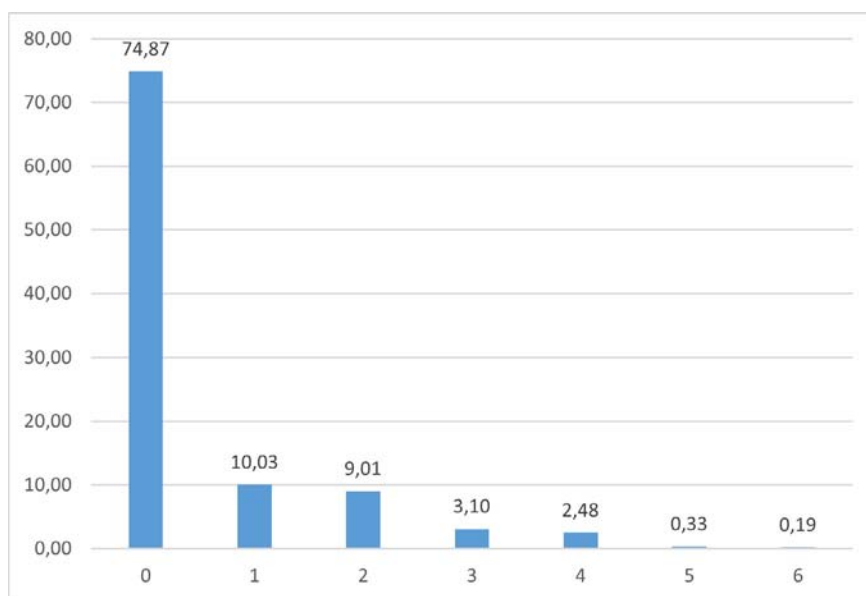
3. Occurrence of Words in Word-Formation Chains and Their Connotative Meanings

3.1 Occurrence of Words in Word-Formation Chains¹¹

As mentioned in the introduction, the multistage word-formation methodology seems to be particularly useful for analyzing material in all three mentioned phases. Based on digitized data from *Besednodružinski slovar slovenskega jezika: poskusni zvezek za iztočnice na B* (Slovenian Word-Family Dictionary: Trial Volume for Headwords Starting with B, 2004) by Irena Stramljič Breznik, which was prepared by Tomaž Erjavec,¹² a statistical analysis was conducted to determine how the word-formation stage affects the occurrence of words in the Gigafida 2.0 and metaFida corpora.

The analysis was initially conducted for all 11,136 derivatives in the word-family dictionary across both corpora. Subsequently, it was performed separately according to the part-of-speech (POS) classification of the simplex, specifically for noun, adjective, verb, and adverb simplexes (there are too few interjection and particle simplexes for the data to be statistically significant). This separation is justified by the fact that the POS classification of the simplex exhibits specific characteristics of the word-formation mechanism (cf. Skarżyński, 1999; 2003; Kern, 2017).

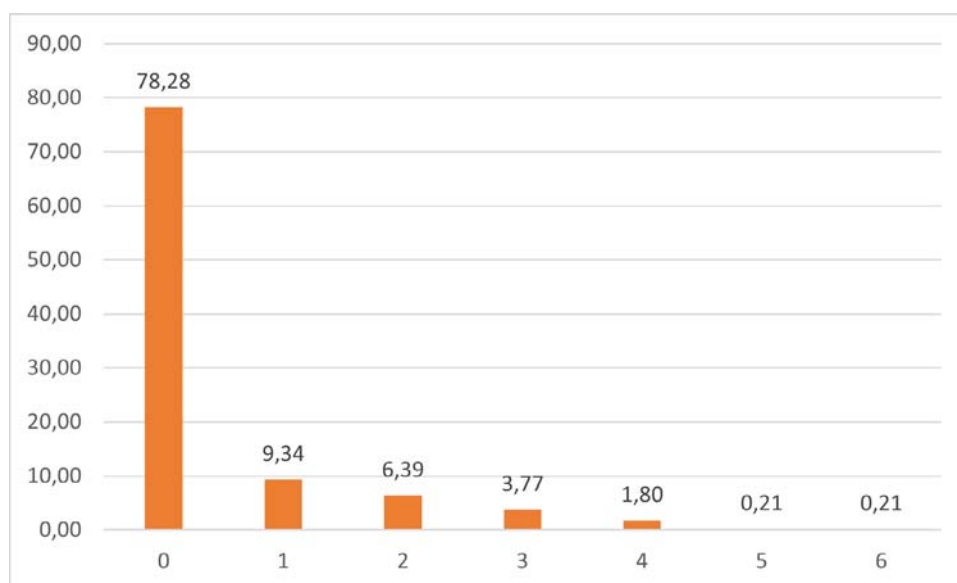
The analysis showed that the Gigafida 2.0 corpus (Graphs 1 and 3) and the metaFida corpus (Graphs 2 and 4) exhibit very similar data. In the *Slovenian Word-Family Dictionary*, the sixth word-formation stage is the highest, meaning that there is no examples of a word-formation chain with more than six stages. The data indicate that the frequency of first-stage derivatives relative to simplex words drops by a factor of 7.5. The occurrence of secondary derivatives relative to first-stage derivatives drops by one-tenth, and a larger drop occurs between the second and third word-formation stage, nearly threefold.



Graph 1: Occurrence of words (in %) according to the word-formation stage in the Gigafida 2.0 corpus

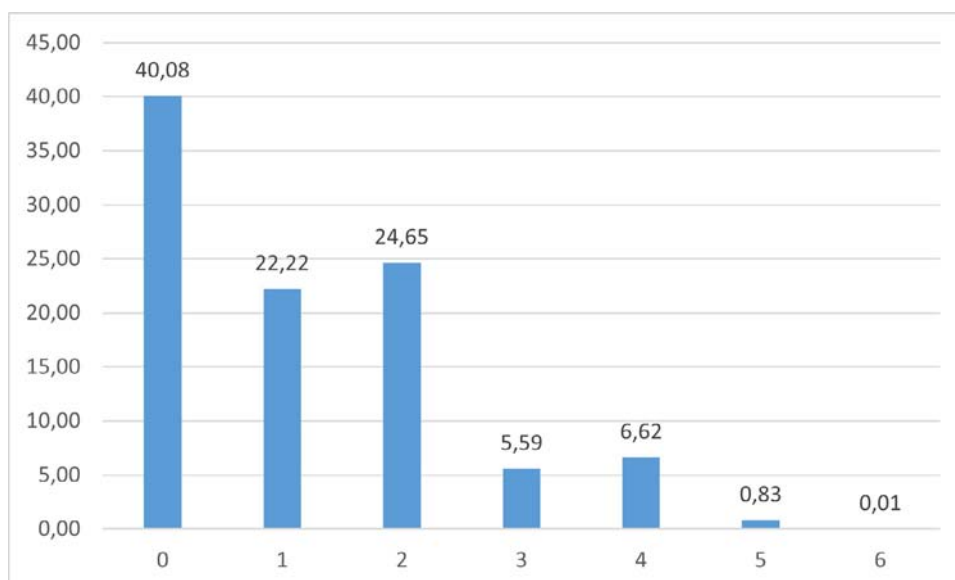
¹¹ The statistical data were prepared by Lazar Pavić.

¹² You can read more about the description of how the word-formation chains were detected in Erjavec et al., 2023.



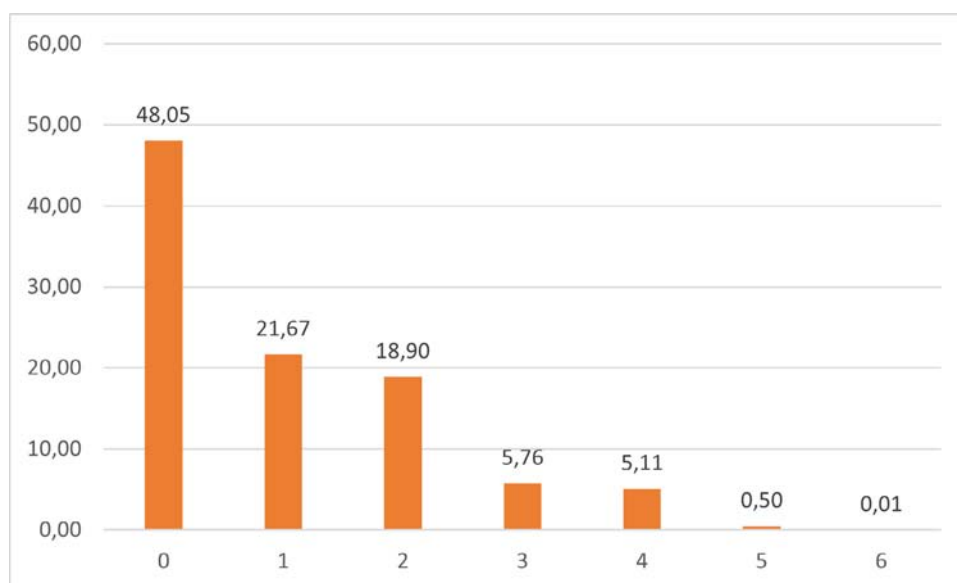
Graph 2: Occurrence of words (in %) according to the word-formation stage in the metaFida corpus

A similar pattern is also observed for nominal and adjectival multistage derivatives, although the data for verbal multistage derivatives differ slightly, and so they are highlighted separately (Graphs 3 and 4). In this group, the frequency of first-stage derivatives decreases by a factor of 1.8, but the frequency of second-stage derivatives, compared to first-stage derivatives, does not decrease but increases by one-tenth. However, a significant drop occurs between the second and third word-formation stages, specifically by a factor of 4.4, whereas the frequency of fourth-stage derivatives increases again.



Graph 3: Occurrence of verbs (in %) according to the word-formation stage in the Gigafida 2.0 corpus

The data from the metaFida corpus do not confirm the frequency increase of second- and fourth-stage derivatives. Instead, the frequency decreases with a similar exponent as in the graph for all words from the word-family dictionary.



Graph 4: Occurrence of verbs (in %) according to the word-formation stage in the metaFida corpus

The data provided are definitely informative and should be considered when setting criteria for including words in dictionaries. In particular, strategies need to be defined for handling higher-stage derivatives, specifically in terms of how much the fact that words belong to the same word family affects the reduction of the frequency criterion and to what extent this reduction is reasonable. It is not possible to provide universal guidelines for determining these thresholds for inclusion in the dictionary; even within a single dictionary, issues often need to be resolved on a case-by-case basis.

3.2 Connotative Meanings of Words in Word-Formation Chains

Previous research on connotative meaning within multistage derivatives (examining verbs of sensory perception; Kern 2018) has shown that both inherent and adherent connotative meanings occur among multistage derivatives. It has also been confirmed for Slovenian that the following depend on the (non)markedness of the word-formation base: 1) the size of the word-formation groups, 2) the number of word-formation models or the absence of certain types of word-formation models in stylistically marked word-formation bases, and 3) the length of word-formation chains. Word-formation groups with a neutral base are more branched than those with connotative meanings in Slovenian as well; word-formation groups with an expressive simplex motivate more multistage derivatives compared to all other bases. In word-formation families with connotative word-formation bases, connotativity is largely retained, with exceptions mainly in some primary verbal derivatives. Connotativity in word-formation groups in which the simplex does not have a connotative meaning primarily appears in first-stage compounds with interfixes (e. g., *drznogled* ‘bold-looking’, *krivogled* ‘cross-eyed’, *zvezdogled* ‘astrologer/astronomer’), (usually) second-stage nominal derivatives with zero suffixes (e. g., *dogled* ‘visibility’, *vpogled* ‘insight’, *oprez* ‘caution’), and higher-stage nominal derivatives (from the third stage onward) resulting from lexicalization (e. g., *pregledanec* ‘examinee’, *spogledljivec* ‘a flirtatious person’). There is less connotativity in word-formation

groups with nominal simplexes. It mainly appears at the second stage, mostly in verbs with prefixal suffixes, which are usually not productive for word formation. It is important to be consistent in dictionaries when providing dictionary qualifiers within word-formation chains. To achieve this goal, it is necessary to review the qualification of all derivatives in the word-formation chains.

4. Conclusions

In compiling dictionaries, considering word formation is essential for ensuring comprehensive and coherent entries. This study highlights the significance of multistage word formation in various phases of dictionary compilation, such as selecting criteria for including words, uniformity in dictionary definitions of related headwords, and consistency in marking stylistically significant senses.

The analysis conducted on the word-family dictionary and corpora, Gigafida 2.0 and metaFida, demonstrates the impact of the word-formation stage on word frequency, which is very informative in compiling dictionaries. The findings indicate a significant decrease in the frequency of derivatives as the formation stage increases, with some variations observed across verbal multistage derivatives. Notably, whereas noun and adjective derivatives follow a consistent pattern, verb derivatives exhibit unique trends, particularly in the second and fourth stages of formation. The data provided should be considered when setting criteria for including words in dictionaries. In particular, strategies need to be defined for handling higher-stage derivatives, specifically in terms of how much the fact that words belong to the same word family affects the reduction of the frequency criterion and to what extent this reduction is reasonable.

Furthermore, previous research on the connotativity of multistage derivatives confirms the presence of both inherent and adherent connotative meanings. The markedness of the word-formation base influences various aspects, such as the size of word-formation groups, the number of formation models, and the length of formation chains. It is important to be consistent in dictionaries when providing dictionary qualifiers within word-formation chains.

In conclusion, integrating word-formation data into dictionary entries enhances the lexicographic work, providing users with a richer and more accurate linguistic resource. Future research should continue to explore the nuances of word formation in Slovenian and other Slavic languages, contributing to the development of innovative and comprehensive dictionaries.

Acknowledgements

This article was written as part of the project Formant Combinatorics in Slovenian (J6-3134), funded by the Slovenian Research and Innovation Agency (ARIS). ARIS funding was also provided through the core program The Slovenian Language in Synchronic and Diachronic Development (P6-0038).

References

- August, G. (1993). New Trends in the Research on Word-Family Dictionaries. *Studia Anglica Posnaniensis*, 25–27, 183–197.
- Burkacka, I. (2012). *Kombinatoryka sufiksalna w polskiej derywacji odrzeczownikowej*. Warsaw: Wydział Polonistyki Uniwersytetu Warszawskiego.
- Čibej, J. (2021). Strojno luščenje medbesednih povezav v oblikoslovnem leksikonu Sloleks 2.0. In Š. Arhar Holdt (Ed.), *Nova slovnica sodobne standardne slovenščine: viri in metode* (1st ed., pp. 125–159). Znanstvena založba Filozofske fakultete. <https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/325/477/7313-1>.
- Dobrovoljc, K., Krek, S., & Erjavec, T. (2017). The Sloleks Morphological Lexicon and its Future Development. In *Dictionary of Modern Slovene: Problems and Solutions*. Znanstvena založba Filozofske fakultete. <https://doi.org/10.4312/9789612379131>.
- Erjavec, T., Pranjić, M., Pelicon, A., Kern, B., Stramljič Breznik, I., & Pollak, S. (2023). Automating derivational morphology for Slovenian. In M. Medved et al. (Ed.), *eLex 2023 : electronic lexicography in the 21st century (eLex 2023) : proceedings of the eLex 2023 conference: [Brno], 27–29 June 2023. Brno: Lexical Computing CZ, 2023* (pp. 449–465). <https://elex.link/elex2023/wp-content/uploads/92.pdf>
- eSSKJ. (2017). Predgovor. https://fran.si/201/esskj-slovar-slovenskega-knjiznega-jezika/datoteke/eSSKJ_Predgovor.pdf.
- Gigafida 2.0: Korpus pisne standardne slovenščine*. viri.cjvt.si/gigafida.
- Gliha Komac, N., et al. (2015). *Koncept novega razlagalnega slovarja slovenskega knjižnega jezika* [electronic source]. Založba ZRC, Znanstvenoraziskovalni center SAZU.
- Gložančev, A. (2012). Novejša slovenska leksika v luči obravnave samostalniških zloženkov v Slovenskem pravopisu 2001. In N. Jakop, & H. Dobrovoljc (Eds.), *Pravopisna stikanja* (pp. 125–150). Založba ZRC, Znanstvenoraziskovalni center SAZU.
- Hudeček, L., & Mihaljević, M. (2016). Tvorbeni podatci. In L. Hudeček, & M. Mihaljević (Eds.), *Hrvatski mrežni rječnik – Mrežnik*. Institut za hrvatski jezik i jezikoslovlje. <http://ihjj.hr/mreznik/uploads/06c8a451ca3cf1d9114ad3744afb4a3e.pdf>
- Kern, B. (2010). Stopenjsko besedotvorje. *Slavistična revija*, 58(3), 335–348.
- Kern, B. (2012). Pisanje skupaj in narazen v Slovarju novejšega besedja slovenskega jezika. In N. Jakop & H. Dobrovoljc (Eds.), *Pravopisna stikanja* (pp. 141–151). Založba ZRC, Znanstvenoraziskovalni center SAZU.
- Kern, B. (2017). *Stopenjsko besedotvorje (Na primeru glagolov čutnega zaznavanja)*. Založba ZRC.
- Kern, B. (2018). Konotativnost z vidika stopenjskega besedotvorja. *Slavistična revija*, 66(2), 130–141.

Krek, S., Kosem, I., & Gantar, P. (2013). *Predlog za izdelavo Slovarja sodobnega slovenskega jezika* [electronic source]. http://www.sssj.si/datoteke/Predlog_SSSJ_v1.1.pdf.

metaFida v1.0. https://www.clarin.si/noske/run.cgi/corp_info?corpname=mfida10&struct_attr_stats=1

Olejniczak, M. (2003). 30 lat słotwórstwa gniazdowego (rys historyczny). In M. Skarżyński (Ed.), *Słotwórstwo gniazdowe: Historia, metoda, zastosowania* (pp. 12–28). Księgarnia Akademicka.

Skarżyński, M. (1999). *Powstanie i rozwój polskiego słotwórstwa opisowego*. Instytut Filologii Polskiej Uniwersytetu Jagiellońskiego.

Skarżyński, M. (2003). Czasowniki jako centra gniazd słotwórczych. In M. Skarżyński (Ed.), *Słotwórstwo gniazdowe: Historia – metoda – zastosowania* (pp. 41–55). Księgarnia Akademicka.

Stramljič Breznik, I. (2004). *Besednodružinski slovar slovenskega jezika: Poskusni zvezek za iztočnice na B*. Slavistično društvo.

Vidovič Muha, A. (2018). *Slovensko skladenjsko besedotvorje*. Znanstvena založba Filozofske fakultete.

Contact information

Boris Kern

ZRC SAZU, Fran Ramovš Institute of the Slovenian Language
University of Nova Gorica, School of Humanities
boris.kern@zrc-sazu.si

Anas Fahad Khan, Ana Salgado, John McCrae, Chamila Liyange, Priya Rani, Rute Costa, Isuri Anuradha, Atul Kr. Ojha, and Francesca Frontini

CULTURAL HERITAGE AND MULTILINGUAL UNDERSTANDING THROUGH LEXICAL ARCHIVES (CHAMUÇA) Portuguese Borrowings in Contemporary Asian Languages

Abstract CHAMUÇA (Cultural HeritAge and Multilingual Understanding through lexiCal Archives) is a pioneering initiative aimed at exploring the impact of the Portuguese language on Asian languages, rooted in the historical exchanges between Portuguese traders, colonists, and diverse Asian cultures. The impact of these interactions extends beyond historical remnants to the modern-day lexicon of Asian languages, which includes a diverse array of Portuguese borrowings, ranging from general vocabulary units to specialised units. We aim to detail the initiative's current status, its goals, and the methodology it employs. Additionally, it will outline the essential steps required for organising and structuring the knowledge embedded within and associated with the borrowings. CHAMUÇA, an innovative open-source resource designed to document and study these Portuguese linguistic contributions, will augment the pool of structured lexical data and support cross-linguistic analysis, using state-of-the-art frameworks such as OntoLex-Lemon and TEI Lex-0 to structure the lexical data. Following FAIR principles – ensuring data is findable, accessible, interoperable, and reusable – CHAMUÇA is poised to contribute to linguistic borrowings, cultural interchange, and the preservation of linguistic heritage. Furthermore, the project will encourage community involvement and scholarly collaboration to evolve and enrich its contents, leveraging collective expertise to illuminate the nuances of language contact phenomena.

Keywords language contact; lexical resource; lexicon; OntoLex; Portuguese language; South Asian languages; TEI

1. Introduction

The current article discusses a newly commenced project CHAMUÇA (Cultural Heritage and Multilingual Understanding through lexiCal Archives) which aims to bring together lexicographers, linguists and other relevant experts to document and analyse the historical influence of the Portuguese language on the languages of Asia, with an initial focus on South Asia. The project will publish its results as a FAIR lexical resource, also named CHAMUÇA, to be based on data from pre-existing language resources, as well as on original scholarly work and crowdsourcing. This resource, which will be released in several versions each under an open licence, is intended to be an open ended one. At the time of

writing, we are working on releasing the first version of CHAMUÇA for a limited number of languages. CHAMUÇA, the language resource, will be made available in two open formats, utilising pre-existing FAIR vocabularies whenever possible. More specifically, it will be published in XML using TEI Lex-0¹, a customisation of the Guidelines for Electronic Text Encoding and Interchange² (TEI-XML) and as Linked Open Data (LOD) using the OntoLex-Lemon³ vocabulary. Each version of CHAMUÇA will offer a comprehensive lexical archive of Portuguese borrowings in each of the target languages that it covers, such as Urdu/Hindi, Sinhala, Tamil, Gujarati, and Bengali.

The rest of the paper is organised as follows. Section 2 gives some historical background on the influence of Portuguese in Asia, with a concentration on the linguistic aspect of this influence. Section 3 shifts to the CHAMUÇA project, detailing its development, focusing on data compilation and the use of conceptual models and guidelines, such as OntoLex-Lemon and TEI Lex-0, for structuring a lexical archive. Finally, Section 4 highlights future work and presents concluding remarks.

2. Legacy of Portuguese in Asian Languages

2.1 Historical Background

The influence of the Portuguese language in Asia began with the so-called Age of Discovery (*Era dos Descobrimentos* in Portuguese) in the 15th and 16th centuries, a period of unprecedented maritime expansion for what was then the Kingdom of Portugal. In particular, during that time several trading posts were established on the Malabar coast, including locations such as Cochin, Cannanore, Quilon, Cranganor, Tanor, and Calicut, giving the Portuguese their first foothold in the continent. Over the next few centuries, and thanks to extensive contact with numerous Asian cultures through trade and colonisation, the Portuguese language was able to leave behind a profound linguistic legacy in the whole of Asia. As Cardoso (2016, p. 71) states, “A língua portuguesa encaixou-se na região asiática ao ponto de se converter em importante língua franca de comércio e diplomacia” [The Portuguese language became deeply rooted in the Asian region to the extent that it evolved into an important lingua franca for commerce and diplomacy]. Despite subsequent, largely successful, moves towards challenging Portuguese dominance in the region by other colonial powers such as the Dutch, French and English, Portugal maintained a presence in colonies such as Goa, Daman, Diu, East Timor, and Macau well into the 20th century.

Today, traces of this influence can still be seen in religious practices and various local cuisines as well as in linguistic borrowings, including toponyms, and, more broadly speaking, the linguistic identities of numerous different Asian communities.

¹ <https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html>

² <https://tei-c.org/guidelines/>

³ <https://www.w3.org/2016/05/ontolex/>

The Portuguese language, once an instrument of trade and empire-building, has inadvertently sown seeds in the rich linguistic soil of Asia, cultivating a lexical heritage that continues to exert a strong influence on modern vernaculars. This influence can be profound, often surfacing in unexpected cultural contexts and leaving behind a rich, indelible legacy that the CHAMUÇA project aims to explore and document. By identifying words of Portuguese origin in Asian languages, the scope of the CHAMUÇA project, we not only document the spread of the Portuguese language in a more global context but also facilitate the study of how these languages adapt and incorporate foreign elements over time.

2.2 Review of Past Literature and Contemporary Scholarship

Cardoso (2016) offers an essential entry point into the scholarly literature on this topic, one that has historical roots stretching back centuries. As well as containing an excellent overview of previous studies on the diffusion of Portuguese throughout Asia and the wider Pacific region from the 16th century onwards, it also details the development of Asian Portuguese creoles. Indeed, the article covers the whole spectrum of linguistic integration, ranging from languages with a modest array of Portuguese-derived words to those which are richly studded with borrowings from Portuguese, to the aforementioned creoles. Cardoso's study also highlights the significance of early Portuguese contact on the Malabar coast and the subsequent linguistic impacts this has had. In terms of historical scholarship, Sebastião Rodolfo Dalgado's pioneering work (Dalgado, 1913) continues to be a cornerstone in this field. In particular, his lexicon of Portuguese borrowings in Asian languages remains an essential lexicographic and scholarly reference which continues to inform and inspire ongoing research, including the CHAMUÇA project.

2.3 Overview of Portuguese Borrowings in South Asian Languages

In this first, preparatory, phase of CHAMUÇA, and given the expertise and interests of the initial participants to the project, we decided to focus on the Portuguese influence in South Asia and South-Eastern Asia, an area covering modern-day countries such as Bangladesh, India, Indonesia, Malaysia, Nepal, Pakistan, and Sri Lanka. This part of the world is rich in linguistic diversity and home to hundreds of languages belonging to the Indo-European, Dravidian, Sino-Tibetan, and Austroasiatic language families. Utilising the framework provided by Cardoso (2016) (which in turn references Dalgado's (1913) study) we list the initial languages covered⁴ by the CHAMUÇA project below (Table 1) with their respective language codes and the number of borrowings for each language listed as given in Dalgado's 1913 lexicon.⁵ In CHAMUÇA we would like to create a lexical resource which makes past research into these borrowings accessible by publishing them in the form of a collection of digital lexicons.

⁴ Note that these are only a subset of the languages covered in Dalgado's original lexicon.

⁵ The current number will presumably be less, although there is the possibility that some newer borrowings such as 'piri-piri' or 'samba' have since then entered these languages, especially via English.

Table 1: Portuguese linguistic borrowings

Languages	Language Code (ISO 639)	Number of Borrowings (Dalgado 1913; Cardoso 2016)
Bengali	bn	173
Hindi/Urdu	hi/ur	53/107
Gujarati	gu	105
Konkani	kok	1768
Malay	ms	431
Marathi	mr	116
Malayalam	ml	127
Sinhala	si	208
Tamil	ta	171
Telugu	te	83

We can make a number of initial observations about the data in the table. Firstly, Konkani (an Indo-Aryan language primarily spoken in the Konkan region, along the western coast of India) shows the highest influence from Portuguese with 1,768 borrowings listed in Dalgado’s work, a figure which is indicative of the deep and prolonged contact with Portuguese culture and language which took place in Konkani. Malay (an Austronesian language spoken in Malaysia, Indonesia, Singapore, Brunei and Thailand) follows with 431 borrowings, evidence of strong maritime trade links and colonial interactions with the Kingdom of Portugal. Hindi/Urdu (widely spoken across the Indian subcontinent) demonstrates a moderate level of Portuguese influence, with 53 and 107 borrowings respectively. Other languages like Bengali (an Indo-Aryan language native to the Bengal region of South Asia), Gujarati (an Indo-Aryan language native to the Indian state of Gujarat), Marathi (an Indo-Aryan language predominantly spoken in the Indian state of Maharashtra), Malayalam (a Dravidian language spoken in the Indian state of Kerala and the union territories of Lakshadweep and Puducherry), Sinhala (an Indo-Aryan language primarily spoken in Sri Lanka), Tamil (a Dravidian language natively spoken by the Tamil people of South Asia), and Telugu (a Dravidian language native to the Indian states of Andhra Pradesh and Telangana) also display a significant number of borrowings, ranging from 83 to 208. This data provides insight into the historical depth of the Portuguese linguistic legacy and helps identify areas for further research in the CHAMUÇA project.

There are many interesting questions to be asked about these borrowings, in particular those looking into the interplay of phonological, morphological, semantic and cultural

factors behind each borrowing. The purpose of CHAMUÇA is to help answer them via a single, easily accessible, machine-actionable language resource. It might be pointed out that a lot of lexical information relevant to this subject – and which could serve to answer the questions alluded to above – is already widely and freely available on sites like *Wiktionary* and *Wikipedia*. However, this information can often be incomplete (e.g., lacking data on corpus frequencies, something which we plan to add in our work), is not always completely reliable, and nor is it often available as structured data (e.g., DBnary, the linked data version of Wiktionary, does not contain structured etymologies and so we cannot extract such borrowings via a SPARQL query). As part of our background work on this project, we tried numerous ChatGPT prompts to ask some of the research questions which we would like to ask of our resource (e.g., ‘list Hindi words which derive from Portuguese’) and as expected the results were not very reliable. On the other hand, a great deal of the scholarly work which exists in this field (as in many others) is distributed across large numbers of articles, scholarly and lexicographic works, and in different languages (largely Portuguese) which are often closed or behind a paywall/unavailable electronically, and rarely as structured language resources.

3. CHAMUÇA as Language Resource

Having been motivated by the considerations described above, we decided that we would make CHAMUÇA an open-source lexical resource published using two existing standards, TEI Lex-0⁶ and OntoLex,⁷ and made available via a Creative Commons Attribution (CC-BY) licence. All of this will help us to ensure that our resource is FAIR by design (although by itself it won’t by itself be sufficient to ensure this, we also intend e.g., to add extensive metadata and publish our resource in a repository). In addition, the use of both TEI Lex-0 and OntoLex means that our resource will be machine-actionable. The project also aims to use crowdsourcing methods through platforms like GitHub to enrich the database with additional lexical units, and detailed lexical information. A dedicated team of linguists, and lexicographers will carefully review and curate all contributions to ensure our resource consistently remains of the highest quality. Native speakers will also contribute to ensure language accuracy. Our vision is for CHAMUÇA to serve as a dynamic, modern counterpart to Dalgado’s 1913 lexicon, and to therefore help in capturing the evolving nature of Portuguese borrowings in Asian languages. This project is not just about preserving knowledge but enhancing how we understand and interact with linguistic heritage in the digital era, potentially making it an important tool for researchers and the general public in navigating cultural and linguistic history and fostering an appreciation of linguistic diversity.

3.1 Building the Lexical Archive

The first step in creating the CHAMUÇA lexical resource is to develop lexicons for the languages listed in Table 1 along with an index lexicon for Portuguese. These

⁶ <https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html>

⁷ <https://www.w3.org/2016/05/ontolex/>

languages were chosen (from among the many we could have worked with) both because they aligned with the linguistic expertise of the researchers participating in the CHAMUÇA project and because each of these languages demonstrates what we consider to be a significant enough level of borrowing from Portuguese (at least in terms of the frequency and common use of the borrowed lexical units if not always in terms of number of borrowings).

This preparatory work for each Asian language lexicon begins by sourcing lexical information from Wiktionary, using a list of candidate lexical units from Dalgado's lexicon (1913) in combination with other lexicographic and scholarly sources (and eventually corpora information). In most cases Wiktionary provides us with basic kinds of linguistic and lexicographic information which we can subsequently modify and enrich – in particular we intend to enrich our lexical entries with relevant cultural, historical, and geographic data where available, leveraging the possibilities offered by linked data to create a rich, multidimensional resource.

The architecture of CHAMUÇA consists of a single lexical resource, which acts as a container for the separate lexicons⁸ for each Asian language covered by the project, all of which are linked to a 'central' Portuguese lexicon. This latter acts as an index for these other lexicons; the architecture of our resource, then, mirrors the structure of Dalgado's 1913 lexicon. We note that the graph-based RDF model of linked data (see Arenas et al., 2009) has proven to be a natural fit for CHAMUÇA, particularly when paired with the graph traversal expressivity of the SPARQL query language.

We have tried to ensure CHAMUÇA's usefulness and relevance by guiding its development via 'competency questions' – specific inquiries that users might potentially ask of our resource via SPARQL queries. These questions range from questions about the domains Portuguese borrowings belong to, to others which concern the specific historical and cultural circumstances of contact for each language. In addition to these kinds of questions, we want our resource to be able to respond to questions respecting the preservation and adaptation of phonological, grammatical, and semantic features, such as gender, across different languages. At a minimum, for each lexical entry in each Asian language lexicon, we capture the following:

- Standard information associated with lexical entries: including the lemma of the entry, its part of speech, transliteration, pronunciation, gender, sense definition, links to related resources including Wiktionary and Wordnet (for the relevant language where available), different morphological variants of the form and other grammatical information
- Etymological information including an original hypothesised Portuguese etymon or etymons and notes
- In the case of Hindi, the equivalent (it exists) in Urdu and vice versa

⁸ In this we were inspired by the definition of the class Lexical Resource in the LMF standard (ISO 24613-1:2019).

- A flag if the word is found in Dalgado's original 1913 lexicon
- One or more domain label tags, see the next section.

At the time of writing, as a test, we have completed the first iteration of our lexicons for Portuguese, Hindi, and Urdu. More information on how these lexicons are structured and their modelling in RDF using OntoLex-Lemon can be found in (Khan et. al, 2024).

3.2 Knowledge Organisation

A pivotal aspect of the construction of the CHAMUÇA lexical resource is the organisation of knowledge to capture the complexity and richness of the original linguistic phenomena being described while ensuring the resource's accessibility and practical utility. To this end, we have adopted domain labelling to categorise and delineate the contexts in which Portuguese borrowings can be found in different Asian languages. Domain labelling foresees the classification of each lexical borrowing into specific categories based on usage and semantic considerations. Note that, by a domain label, we mean a 'marker which identifies the specialised field of knowledge in which a lexical unit is mainly used' (Salgado, Costa, and Tasovac, 2019). A manual approach is chosen to ensure precise and context-aware categorisation, capturing subtle nuances that automated systems might overlook. This organisation method facilitates a user's navigation through the resource, helping them understand the integration of Portuguese lexical items into the relevant cultural and functional contexts. For example, the Hindi and Urdu words for 'church' are both borrowed from the Portuguese *igreja*, गरिजा and اچرگ (both with the pronunciation girajā), respectively and are categorised under Religion, underscoring the specific cultural significance of this borrowing. The influence of the Portuguese language is manifest in its introduction of new lexical units across multiple domains as noted by Dalgado (1913, pp. XXVII-XXIX). These impacts are organised into several principal domain labels such as the following:

Religion. This domain label includes units related to Christian practices and concepts introduced by Portuguese missionaries. For instance, the Hindi and Urdu words for 'cross', (क्रूस Hi., سورك Ur.) are derived from *cruz*, and those for '(Christian) priest' (पादरी Hi., پادری Ur.) from *padre*. This labelling not only aids in the study of linguistic influence in the context of religious conversion but also facilitates the tracing of how these terms were adopted and adapted by other cultures.

- **Botany.** The domain covers units referring to plant species introduced across the Asian continent which retained their original Portuguese names. This is the case for the Hindi and Urdu words for pineapple, (अननस Hi., ساننا Ur.), which both derive from *ananás*, and those for cabbage (गोभी Hi., یهبوگ Ur.), from *couve*, at least according to one hypothesis. This focus allows for a detailed study of how these plant names propagated across different regions along with the plants themselves.

- **Culinary.** The significant impact of Portuguese influence on Indian cuisine is evident in a number of culinary terms. It was responsible for introducing new ingredients and products, including chilli peppers, which revolutionised Indian cooking. An exemplary term is *balti*, originally from the Portuguese word *balde* ‘bucket’, now synonymous with a style of Punjabi cuisine popular in the UK. This reflects not only the culinary practices but also the journey of a word across cultures.
- **Clothing.** Lexical borrowings demonstrate how Portuguese fashion influenced local attire, filling gaps and adding diversity to the clothing vocabulary in Asian languages.
- **Cultural:** This domain indicates items of everyday use and material culture, such as the words for ‘cabinet’ in Hindi and Urdu which are borrowed from *armário* (अलमारी Hi., یراملأ Ur.), the word for ‘bucket’ from *balde* (बाल्टी Hi., ٲٹللاب Ur.)⁹, and pistol from *pistola* (पस्तौल Hi., لوطسپ Ur.). Analysing these lexical units helps in understanding how the Portuguese language influenced daily life and material culture through trade and colonial interactions.

Through these organised domains, CHAMUÇA helps to contribute towards a general understanding of cultural linguistics.

4. Final Considerations and Future Work

The CHAMUÇA project aims to describe and assist into research on the impact of Portuguese borrowings into Asian languages, helping to provide data on the complex dynamics of language contact and cultural exchange. By employing advanced lexical frameworks and adhering to FAIR principles, we aim to ensure that the data is not only accessible but also interoperable across different platforms.

While the initial phase focused on South Asian (in addition to one Southeast Asian language), future expansions will include East and Southeast Asian languages, broadening the scope of our research and making the archive more inclusive. We aim to collaborate with linguists, historians, and cultural scholars to explore the broader implications of linguistic borrowings. These collaborations will help contextualise the lexical data within larger narratives of historical interactions and cultural transformations.

Ensuring the long-term sustainability of CHAMUÇA is a priority. We will seek additional funding sources and partnerships to maintain and expand the archive.

References

Arenas, M., Gutierrez, C., & Pérez, J. (2009). Foundations of RDF databases. In *Reasoning Web International Summer School* (pp. 158–204). Springer Berlin Heidelberg.

⁹ Note that in cases where a borrowing has more than one sense we can either tag at the sense level or the level of the whole entry. In the case of the borrowing of *balde* for instance we will tag at the sense level.

Cardoso, H. (2016). O português em contacto na Ásia e no Pacífico. In A. M. Martins, & E. Carrilho (Eds.), *Manual de Linguística Portuguesa* (pp. 68–97). Berlin, Mouton de Gruyter.

Cimiano, P., McCrae, J. P., & Buitelaar, P. (2016). *Lexicon Model for Ontologies: Community Report, Technical Report*. W3C Ontology-Lexicon Community Group. <https://www.w3.org/2016/05/ontolex/>

Dalgado, S. R. (1913). *Influência do vocabulário português em línguas asiáticas (abrangendo cerca de cinquenta idiomas)*. Imprensa da Universidade.

Khan, F., Salgado, A., Anuradha, I., Costa, R., Liyanage, C., McCrae, J. P., Ojha, A. K., Rani, P., & Frontini, F. (2024). CHAMUÇA: Towards a Linked Data Language Resource of Portuguese Borrowings in Asian Languages. In *Proceedings of the 9th Workshop on Linked Data in Linguistics @ LREC-COLING 2024* (pp. 44–48). ELRA and ICCL.

Salgado, A., Costa, R., & Tasovac, T. (2019). Improving the consistency of usage labelling in dictionaries with TEI Lex-0. Lexicography. *Journal of ASIALEX*, 6(2), 133–156. doi:10.1007/s40607-019-00061-x

Tasovac, T., Romary, L. et al. (2018). *TEI Lex-0: A baseline encoding for lexicographic data. Version 0.9.1*. DARIAH Working Group on Lexical Resources. <https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html>, last accessed 26 January 2024.

Acknowledgements

The research of Ana Salgado and Rute Costa is supported by the Portuguese national funding through the FCT – Portuguese Foundation for Science and Technology, I.P. as part of the project UIDB/LIN/03213/2020;10.54499/UIDB/03213/2020 and UIDP/LIN/03213/2020;10.54499/UIDP/03213/2020 – Linguistics Research Centre of NOVA University Lisbon (CLUNL). John P. McCrae, Atul Kr. Ojha and Priya Rani would like to acknowledge the support of the Science Foundation Ireland (SFI) as part of Grant Number SFI/12/RC/2289_P2 Insight_2, Insight SFI Research Centre for Data Analytics.

Contact information

Anas Fahad Khan

CNR-ILC, Italy
fahad.khan@ilc.cnr.it

Ana Salgado

NOVA University Lisbon, Portugal
anasalgado@fcsh.unl.pt

John McCrae

University of Galway, Ireland
john.mccrae@universityofgalway.ie

Chamila Liyange

University of Colombo, Sri Lanka
cml@ucsc.cmb.ac.lk

Priya Rani

University of Galway, Ireland
priya.rani@insight-centre.org

Rute Costa

NOVA University Lisbon, Portugal
rute.costa@fcsh.unl.pt

Isuri Anuradha

University of Wolverhampton, UK
Isuri.Anuradha@wlv.ac.uk

Atul Kr. Ojha

University of Galway, Ireland
atulkumar@insight-centre.org

Francesca Frontini

CNR-ILC, Italy
francesca.frontini@ilc.cnr.it

Veronika Kolářová and Jiří Mírovský

LOOKING FOR SENSE IN NONSENSE: VALENCY OF NEGATIVE FORMS OF NOUNS AND ADJECTIVES IN THE NOMVALLEX LEXICON

Abstract NomVallex is a manually annotated valency lexicon of Czech nouns and adjectives that enables research into various language phenomena related to valency, including the comparison of valency properties of affirmative and negative forms of words. This paper presents new developments in the way the lexicon facilitates research into word-level negation, explaining the reasoning behind the proposed lexicographic treatment. Differentiating between direct negation and lexicalized negation, we discuss whether or not the negative forms of words should be listed separately in a valency lexicon. We argue that, while lexicalized negation has to be assigned an individual entry, direct negation of nouns and adjectives should be treated within an entry for affirmative forms. Considering various aspects of word-level negation, including the employment of negative forms of nouns and adjectives in light verb constructions or phrasemes, we describe negation-related attributes applied to the data of the lexicon. As a case study, the facilities provided by the lexicon are used to illustrate use and distribution of negative forms of Czech deadjectival nouns, and their valency properties are compared to those of the corresponding affirmative forms.

Keywords negation; negative form; affirmative form; nouns; adjectives; valency lexicon

1. Introduction

Negation can be expressed using various grammatical processes, especially sentential negation, phrasal negation and word-level negation (Curiel, 2015). Word-level (lexical) negation, e.g., *independence*, affects different word classes in different ways (Pavlovič, 2015), and thus may attract lexicographers' attention.

Kováříková et al. (2012) discuss whether or not the negative forms of words should feature as entries in a monolingual dictionary. Based on corpus data, the authors investigate whether negation of Czech verbs, nouns, adjectives and adverbs can be considered a grammatical category (Kováříková, 2011), and they analyze the coverage of negative forms (N-forms) and affirmative forms (A-forms) of verbs, adjectives and adverbs in different frequency levels. They conclude that, while N-forms of verbs should not be assigned an individual entry in the dictionary, N-forms of nouns, adjectives and adverbs should be listed separately (the two latter provided that they are frequent enough).

In specialized dictionaries, for instance, in valency lexicons (for English, see Herbst et al., 2004; for Czech, see Hajič et al., 2003, Svozilová et al., 2005), N-forms of nouns and adjectives are usually accorded a separate entry, too.

In this paper, we deal with the treatment of negation in NomVallex, a valency lexicon of Czech nouns and adjectives (Kolářová & Vernerová, 2022, Section 4). The aims of the study are twofold: (i) theoretical research into various aspects of lexical negation with respect to the meaning and valency of nouns and adjectives; (ii) description of the lexicographic treatment of negation in NomVallex. First, differentiating between direct negation and lexicalized negation, we outline the status of word-level negation in Czech word formation (Section 2). Further, we specify aspects that should be considered when treating negation in a valency lexicon (Section 3). Then we introduce the NomVallex lexicon (Section 4.1) and describe how negation is treated in the lexicon data (Section 4.2). Finally, as a case study, preliminary results of the annotation are illustrated through the case of negative forms of Czech deadjectival nouns and their valency (Section 5).

2. Word-Level Negation and Word Formation

From the word-formation point of view, word-level negation can be seen as the combination of a negative formant (negator) and an affirmative base (Pavlovič, 2015). European languages usually express word-level negation by affixation (Zimmer, 1964, Curiel, 2015, van Son et al. 2016). In Slavic languages, the universal exponent of negation is the morpheme *ne-* (Pavlovič, 2015; e.g., Cz. *nezávislost* ‘independence’), used with autosemantic word classes, i.e., nouns, adjectives, verbs and adverbs.

Focusing on Czech, when the negative prefix *ne-* is combined with verbs, it typically only denies the affirmative content, i.e., the original predication, without specifying new features (Lotko, 1975; cf. *záviset* ‘depend’ vs. *nezáviset* ‘not to depend’); therefore, N-forms of verbs are not treated in any specific way in Czech lexicons. In contrast, with nouns, adjectives and adverbs, the situation is more complex: in addition to direct (mere) negation, e.g., *(ne)závislost* ‘(in)dependence’, *(ne)šťastný* ‘(un)happy’ (Joshi, 2020, p. 86), the prefix *ne-* may lead to a semantic shift (Rechzieglová, 1995, Pavlovič, 2015, Kolářová, 2020; cf. *smysl* ‘sense’ vs. *nesmysl* ‘nonsense’, *ohrožený* ‘threatened’ vs. *neohrožený* ‘fearless’, *patrně* ‘apparently’ vs. *nepatrně* ‘slightly’), referred to here as lexicalized negation. Moreover, N-forms can be used as negative occasional words, for instance in disjunctive syntactic constructions of the type *Zákaz – nezákaz, nalej!* ‘Forbidden or not, pour!’ (Pavlovič, 2015, p. 1365). In order to express meanings slightly different from direct negation, the prefix *bez-* ‘without’ is used in Czech (e.g., *bezdlužnost* ‘state of being without debts’, *bezmocný* ‘powerless’).

The possibility of direct negation in Czech differs according to individual word classes: while all verbs and most adjectives can be negated (Štícha et al., 2013, Pavlovič, 2015, Rusínová & Osolobě, 2016, Šimandl et al., 2016), when it comes to nouns, usually only abstract nouns allow direct negation (Štícha et al., 2013). Furthermore, abstract nouns differ in their use of direct negation depending on their derivational type and semantic category (cf. action, e.g., *(ne)komunikace* ‘non-communication’, and substance, e.g., *komunikace* ‘road’). As for deverbal nouns, only so-called stem nouns (i.e., nouns that in Czech end in *-ní/-tí* and contain a theme suffix of the base verb, e.g., *namítání* – *namítnutí* ‘raising objections’) can be systematically negated if they denote action. In

contrast, so-called root nouns (derived from verbs by various suffixes, including the zero suffix, but not containing a theme suffix, e.g., *námitka* ‘objection’) can be negated only in isolated cases (Dvořák, 2016a, Dvořák, 2016b, Kolářová, 2020). Deadjectival nouns (e.g., *závislost* ‘dependence’) derived from adjectives using N-forms usually allow direct negation, too (Kolářová, 2020; Section 5).

3. Negation and Valency

Negative prefixes in the strict sense do not normally affect the argument structure (valency) of the base (Curiel, 2015); however, exceptions exist, both with nouns and adjectives (Haugen, 2013, Kolářová, 2020). Moreover, N-forms can be employed in light verb constructions (Kettnerová et al., 2018) or phrasemes, e.g., *přinést neklid* ‘bring uneasiness’, *plácet nesmysly* ‘talk nonsense’.

When treating negation in a valency lexicon, the following aspects should be considered:

- (i) N-forms that have undergone a semantic shift should undoubtedly be assigned an individual entry (lexeme); the difference in meaning can be accompanied by difference in valency, e.g., *nedůtklivost* ‘touchiness’ has two complements (e.g., *Petrova nedůtklivost vůči kritice* ‘Peter’s touchiness on criticism’) whereas *důtklivost* ‘urgency’ only has one (e.g., *důtklivost jeho výkladu* ‘the urgency of his presentation’);
- (ii) The N-form is the direct negation of the A-form:
 - (ii-a) Typically, A-forms and N-forms have the same valency, and thus separate entries for both A-form and N-form may be regarded as redundant, just as in the case of verbs;
 - (ii-b) Rarely, there is a difference in valency between A-forms and N-forms, cf. Eng. *dependence on* and *independence of / from*, which should then be treated in separate entries;
 - (ii-c) Commonness: N-forms and their valency are usually less common (frequent) than A-forms (e.g., *nevděčný za něco* ‘ungrateful for sth’), however, some N-forms considerably outnumber A-forms (e.g., *nepostradatelný pro někoho* ‘indispensable to sb’).
- (iii) There are cases when only one of the forms (either A-form or N-form) can be a part of a light verb construction or a phraseme, e.g., Cz. *vyhlásit nezávislost* ‘declare independence’ vs. **vyhlásit závislost* ‘declare dependence’, Eng. *make sense to / for sb* vs. *make nonsense of sth*.

4. Lexicographic Treatment of Negation in NomVallex

4.1 Organization of the NomVallex Lexicon

NomVallex is a manually created valency lexicon of Czech nouns and adjectives (Kolářová & Vernerová, 2022), adopting the theoretical framework of Functional Generative Description as its theoretical basis. Its newest version, NomVallex 2.0 (Kolářová et al., 2022), comprises 1,027 lexical units contained in 570 lexemes.

NomVallex follows in the footsteps of the VALLEX lexicon (a valency lexicon of Czech verbs created within the same theoretical framework; Lopatková et al., 2022) and adopts its annotation scheme (see Figure 1). The lexicon entry in both lexicons is called a lexeme; it is an abstract unit associating lexical forms with their lexical units (LUs; Cruse, 1986), i.e., word senses. The entries capture all lexical meanings (senses) of the words included in the lexicon; the lexical units are given their identifier (id), consisting of the string *blu* (i.e., an abbreviation for *basic lexical unit*), letters *v/n/a* differentiating among verbs, nouns, and adjectives, lemma of the word, and the number of the lexical unit (cf. the id *blu-n-závislost-1* of the first lexical unit of the lexeme *závislost* ‘dependence’ in Figure 1). The particular lexical meanings are described by synonyms or roughly paraphrased in a gloss (the *synon* attribute). Aspectual (imperfective vs. perfective) counterparts formed by suffixation, such as *vyzývání^{impf}* – *vyzvání^{pf}* ‘appealing’ or *ohrožování^{impf}* – *ohrožení^{pf}* ‘threatened’, are treated within a single lexeme. Nouns or adjectives that do not express aspect are assigned the flag *no-aspect*, e.g., *výzva^{no-aspect}* ‘appeal’.

The valency properties of a lexical unit are captured in a valency frame, consisting of a set of slots, each standing for one valency complement. Each valency complement is assigned a functor (a label marking the relation of the valency complement to its governing word), and a list of forms determining the surface realization of the complement. The following types of complements may be a part of valency frames: obligatory or optional actants (i.e., ACTor, PATient, ADDRessee, EFFECT, and ORIGIN, e.g., *Petrova_{ACT} výzva k pomoci_{PAT}* ‘Peter’s appeal for help’, *prodejný mládeži_{ADDR}* ‘marketable to young people’), and obligatory free modifications, especially those with the meaning of direction (e.g., *muž povolaný do armády_{DIR3}* ‘a man drafted into the army’). In NomVallex, the valency properties of a lexical unit are documented in the example-rich attribute by examples from the Czech National Corpus (two corpora are used, namely SYNv12 (Křen et al., 2023) and Araneum Bohemicum Maximum (Benko, 2015)). The valency frame of the first lexical unit of the noun *závislost* ‘dependence’ and examples documenting it are exemplified in Figure 1.

Extending the basic VALLEX annotation scheme, NomVallex employs new attributes and facilities in order to research phenomena related to part-of-speech specificity of the nouns and adjectives covered (Kolářová et al., 2023). In NomVallex, each lexical unit of an adjective or a noun is assigned a label indicating its derivational type (filled in the attribute type), see Table 1. The label provides information on both part-of-speech membership of the LU (whether it is a noun (N) or an adjective (A)) and its derivational base (whether it is deverbal (DV), deadjectival (DA), denominal (DN) or

primary (P)); for more details on the labels, see Kolářová et al., 2023, e.g., the noun *závislost* ‘dependence’ is given the label N-DA-3-lost.¹ The structure of the NomVallex 2.0 lexicon, reflecting the basic derivational categories captured in the lexicon, is presented in Table 1.

Table 1: The structure of the NomVallex 2.0 lexicon

Part-of-speech category	Derivational category	Derivational type	Lexical units	Lexemes
Nouns	deverbal	N-DV-...	505	248
	deadjectival	N-DA-...	247	158
Adjectives	deverbal	A-DV-...	179	116
	denominal	A-DN-...	28	14
	deadjectival	A-DA-...	6	6
	primary	A-P	62	28
Total			1,027	570

To enable analysis of the relationship between the valency behavior of base words and their derivatives, lexical units of nouns and adjectives in NomVallex are linked to their respective base lexical units (contained either in NomVallex itself or, in the case of verbs, in the VALLEX lexicon). Derivationally related lexical units of nouns and adjectives are linked to each other (or to their base verbs in VALLEX) by means of two attributes which keep both directions, namely (i) the attribute *derivedFrom* (it provides a link from a particular LU to its base LU), and (ii) the attribute *derivedLUs* (being dynamically generated according to the latest data, it captures links to all LUs derived from the base LU).

The treatment of negative forms of nouns and adjectives presented in this paper is a new development in the lexicon data that will feature in the future published version (in 2024).

4.2 Negation-Related Attributes in NomVallex

As a manually annotated lexicographic resource, the NomVallex lexicon aspires to treat negation of nouns and adjectives as accurately as possible. All lexical units of nouns and adjectives contained in NomVallex are examined with respect to whether or not they can be used in N-forms expressing direct negation of the A-forms. The N-forms are searched for in the two corpora used for documenting the A-forms, namely in SYNv12 and Araneum Bohemicum Maximum. An annotator then looks through the occurrences found and manually annotates them in the data of the lexicon,

¹ In particular, the label N-DA-3-lost indicates a noun derived from an adjective (i.e., *závislost* ‘dependence’ < *závislý* ‘dependent’) that is derived from an active participle (this verbal form being marked here by number 3; i.e., *závislý* ‘dependent’ < *závisel* ‘depended’). Moreover, as the mnemonic aid, the label contains the string in which this derivational type of noun ends, namely *-lost*.

differentiating between particular lexical units (senses) of the N-forms, if necessary. If a word among the N-forms turns out to have its own (lexicalized) meaning, then it is assigned an individual entry (lexeme; e.g., *nezvyklý* ‘unusual’).²

In order to tackle the issues specified in Section 3, N-forms in NomVallex that directly negate lexical units expressed in A-forms and that do not differ in valency are captured within the entries for A-forms. In extension of the basic annotation scheme described in Section 4.1, the negation-related attributes specified in Table 2 are employed, if relevant (the treatment of the N-form *nezávislost* ‘independence’ is exemplified in Figure 1).

```
* ZÁVISLOST [NomVallex.txt]
-lexeme-affirm-freq: 187612
-lexeme-affirm-i-p-m: 30,08
-lexeme-neg-freq: 98257
-lexeme-neg-i-p-m: 15,75
~ no-aspect: závislost [blu-n-závislost-1]
+ ACT2, pos PATna+6
-synon: odkázanost
-derivedFrom: blu-a-závislý-1
-example-rich: ACT+PAT: Zdejší zásoby závislost Slovenska.ACT na dovozu.PAT plynu z Ruska nezmění. •
  Za své selhání považovali přílišnou závislost dětí.ACT na matce.PAT. • S tím, jak se zvyšuje počet
  vysokoškolských studentů, se prodlužuje období jejich.ACT závislosti na rodičích.PAT. • Vojenských
  alternativ měli poskrovnou, jejich.ACT závislost na Západu.PAT sílíla
-type: N-DA-3-lost
-neg-lemmas: no-aspect: nezávislost
-neg-id: blu-n-závislost-1:neg
-neg-commonness: standard
-neg-type: direct negation
-neg-example: ACT+PAT: To by umožnilo její.ACT nezávislost na politických stranách.PAT. • Předložený
  studijní grant, který by studentovi měsíčně přispíval na studia a zvyšoval tak jeho.ACT nezávislost na
  rodičích.PAT, se pro změnu neopírá o žádnou finanční rozvahu. • Zvýší se také jeho.ACT nezávislost na
  sousedním Rusku.PAT. • Půjčky se tradičně užívají v severských zemích a v Nizozemsku: pro tyto země
  je totiž charakteristická brzká nezávislost dětí.ACT na rodičích.PAT. • V USA těžba plynu a ropy z břidlic
  znamená faktickou nezávislost Spojených států.ACT na importech.PAT.
-neg-lvc: blu-v-vyhlásit-vyhlašovat-4
```

Fig. 1: The N-form *nezávislost* ‘independence’ is captured within the entry for the A-form *závislost* ‘dependence’

First, the attribute *neg-type* specifies whether the N-form represents direct negation or lexicalized negation. The attribute *neg-type* also captures cases of lemmas that do not express word-level negation, differentiating between so-called *negativum tantum* (the term used in Czech terminology for words that start with the string *ne-* but that in present-day Czech have no meaning without the string, e.g., *nedočkávy* ‘impatient’, *nenávist* ‘hatred’) and words considered not to be able to occur in the N-forms for various reasons (the value *inapplicable* is used in this case).

² As NomVallex specializes in non-verbal predicates endowed with valency, the lexicographic treatment of negative forms of Czech nouns and adjectives covers only a small part of Czech lexis. As for nouns, in the SYNv12 corpus there are 8,876 nouns, the lemmas of which start with the string *ne-* (including words that have no relation to negation, such as *nebe* ‘sky’ or *nerv* ‘nerve’), whereas NomVallex only features slightly more than 400 noun lexemes. As for adjectives, the SYNv12 corpus contains 20,207 lemmas of adjectives that have at the 11th position of their morphological tag the label N used for negation, whereas NomVallex captures fewer than 200 adjectival lexemes.

N-forms representing direct negation are assigned the particular negative lemma(s) in the ‘neg-lemmas’ attribute. With lexicalized negation, negativum tantum or inapplicable N-form(s), the value of the ‘neg-lemmas’ attribute is ‘NA’. If a lexical unit has no hit for its N-form in the corpora but it could theoretically be used, the value of its ‘neg-lemmas’ attribute is the appropriate N-form, and the ‘neg-commonness’ attribute is assigned the value ‘NO example’. Nevertheless, it is sometimes difficult to distinguish clearly between the two cases of no occurrence of the particular N-form.

Table 2: Negation-related attributes of lexical units in the NomVallex lexicon

Attribute	Value	Example
neg-type	direct negation	<i>nezávislost</i> ‘independence’
	lexicalized negation	<i>neohrožený</i> ‘fearless’
	negativum tantum	<i>nenávist</i> ‘hatred’
	inapplicable	<i>nezkouška</i> ‘non-exam’
neg-lemmas	lemma(s) of the N-form	<i>nezávislost</i> ‘independence’
	NA	* <i>nedomněnka</i> ‘unsurmise’
neg-commonness	NO example	<i>neapelování</i> ‘non-appealing’
	rare	<i>nezvědavost</i> ‘incuriosity’
	rare, marked	<i>nepodmínka</i> ‘non-condition’
	standard	<i>nezávislost</i> ‘independence’
	more frequent than the A-form	<i>nepostradatelný</i> ‘indispensable’
neg-example	annotated corpus examples	<i>To by umožnilo její_{ACT} nezávislost na politických stranách_{PAT}.</i> ‘It would enable her independence from political parties.’
neg-id	identifier of the N-form	blu-n-závislost-1:neg
neg-lvc	identifier(s) of light verb(s) with which the N-form forms a predicate	blu-v-vyhlásit-vyhlašovat-4 (cf., <i>vyhlásit</i> ‘declare’ + the N-form <i>nezávislost</i> ‘independence’)

The other values of the ‘neg-commonness’ attribute express the subjective annotator’s evaluation of how common the N-forms are with respect to the corresponding A-forms (primarily examples with valency are taken into account); the values include ‘rare’, ‘rare, marked’, ‘standard’, and ‘more frequent than the A-form’ (see Section 5 for more details). The values of the ‘neg-commonness’ attribute are somewhat vague (i.e., not numerical), but thanks to manual annotation they are able to capture the differences between particular senses of a lexeme.

In addition to the ‘neg-commonness’ attribute, the NomVallex data are also automatically supplemented with attributes providing the absolute and relative corpus frequency of both

affirmative and negative forms of the words included (the SYNv12 corpus is used). However, the corpus frequency (by definition) cannot distinguish between individual lexical units (senses) (e.g., *prodejnost výrobku* ‘marketability of the product’ vs. *prodejnost politika* ‘venality of a politician’), homonyms (e.g., *výklad* ‘explanation’ vs. *výklad* ‘shop window’) or reflexive and non-reflexive lexemes (cf., *zážitek nedorozumění se s někým* ‘experience of not understanding each other’ and *malé nedorozumění* ‘slight misunderstanding’). Therefore, despite being ‘objective’, the information value of corpus frequency is in fact in many cases only approximate. As the frequency data are not relevant to individual lexical units, attributes covering corpus frequency are listed at the level of lexemes, not lexical units (attributes capturing corpus frequency of the affirmative and negative lemmas of the lexeme *závislost* ‘dependency’, i.e., ‘lexeme-affirm-freq’, ‘lexeme-neg-freq’, ‘lexeme-affirm-i-p-m’, and ‘lexeme-neg-i-p-m’, are exemplified in Figure 1).³

The attribute ‘neg-example’ contains manually annotated examples, handpicked from the corpora used; the examples should primarily document the valency properties of the particular N-forms.

In order to tackle the cases when only one of the forms (either the A-form or the N-form) can be a part of a phraseme or a light verb construction (lvc; e.g., *vyhlásit nezávislost* ‘declare independence’), in addition to the A-form’s identifier (e.g., blu-n-závislost-1), also each N-form annotated within a lexical unit together with the A-form is given its own identifier (the ‘neg-id’ attribute, e.g., blu-n-závislost-1:neg). As a result, the particular N-form and the light verb can be interlinked (e.g., *nezávislost* ‘independence’ and the light verb *vyhlásit* ‘declare’). If relevant, the identifier(s) of the light verb(s) with which the N-form forms a predicate are specified in the ‘neg-lvc’ attribute.

5. A Case Study: Negative Forms of Deadjectival Nouns in NomVallex and Their Valency

The manually annotated negation-related attributes (Section 4.2) provide an overall picture of the use of negative forms of nouns and adjectives contained in the NomVallex lexicon. To illustrate the use and distribution of the negative forms, including their valency properties, we focus here on one derivational category annotated in accordance with the guidelines described in Section 4, namely on Czech deadjectival nouns (in total 247 lexical units in 158 lexemes, see Table 1; e.g., *závislost* ‘dependence’ < *závislý* ‘dependent’).

We assume that the values of the ‘neg-commonness’ attribute in particular (described below) can serve as a good indicator of how frequent or common the N-forms are compared to their corresponding A-forms. The distribution of the values assigned in our data to deadjectival nouns is presented in Table 3.

³ Nevertheless, homonyms and reflexive vs. non-reflexive words captured in the lexicon are not provided with the corpus frequency attributes because they would not be relevant in these cases.

Table 3: The ‘neg-commonness’ attribute of lexical units of deadjectival nouns in NomVallex

N-form: direct negation	Value of the attribute	Number of lexical units	
		absolute	relative
No	not present	34	13,8
Yes	standard	82	33,2
	rare	75	30,4
	rare, marked	6	2,4
	more frequent than the A-form	3	1,2
	NO example	47	19,0
Total		247	100%

While nouns in the first category (labeled with the value ‘not present’) clearly differ from typical A-form – N-form pairs and are mentioned here for the sake of completeness only, all the other categories represent cases of direct negation of Czech deadjectival nouns:

- ‘not present’: The lexical units in question are represented by (i) lexicalized N-forms (e.g., *neohroženost* ‘fearlessness’), (ii) nouns that in present-day Czech have no meaning without the string *ne-* (e.g., *nedočkavost* ‘impatience’), (iii) nouns to which the N-form is not relevant for some other reason, for example because their A-form is lexicalized or is employed in a phraseme and thus cannot be negated (e.g., *drahota* used in the phraseme *dělat drahoty* ‘make a fuss’). Their attribute ‘neg-lemmas’ receives the value ‘NA’ (see Section 4.2), and thus the ‘neg-commonness’ attribute is not relevant to them.
- ‘standard’: Being quite frequent, these N-forms use the same valency as their A-form counterparts, such as the N-form *nezávislost* ‘independence’ and the A-form *závislost* ‘dependence’, both amply documented with Actor expressed in the genitive or possessive forms and with Patient in a prepositional phrase introduced by *na* ‘on / from’ (see Figure 1).
- ‘rare’: These N-forms are usually documented in few examples, being considerably less frequent than their corresponding A-forms. The N-forms are only rarely modified by valency complements (e.g., *nehrdost* ‘non-pride’); moreover, usually also some of the complements (or some of their forms) are typically not documented at all, although they probably could be used.
- ‘rare, marked’: These N-forms are not only rare, but the examples found in the corpora used even seem to be ‘marked’ or ‘odd’ (e.g., *nevlastnost* ‘non-property’).
- ‘more frequent than the A-form’: In isolated cases, N-forms far outnumber the corresponding A-forms (e.g., *nedotknutelnost* ‘untouchability’).
- ‘NO example’: These N-forms are not documented in any example in the corpus data. However, they could theoretically be used (e.g., *neuraženost* ‘non-pique’).

The distribution of the values of the ‘neg-commonness’ attribute in our data suggests that the use of N-forms of Czech deadjectival nouns more or less corresponds to tendencies described in previous studies (see e.g., Kolářová, 2020). Most deadjectival nouns seem to use their N-forms for expressing direct negation. However, only about one third of the nouns in our data use their N-forms fairly frequently and with the same valency as the corresponding A-forms (the ‘standard’ N-forms in Table 3). Leaving aside cases of non-expression of valency complements, overall, only slight differences in valency between A-forms and N-forms are attested in our data. The prepositional phrases of complements other than Actor are usually kept with the N-forms, and they are sometimes preferred to prepositionless cases; for instance, while the Patient of the A-form of the noun *škodlivost* ‘harmfulness’ is documented in a prepositionless dative (e.g., *jeho škodlivost lidskému zdraví* ‘its harmfulness to human health’), it is not attested with the N-form *neškodlivost* ‘non-harmfulness’. Instead, both A-form and N-form (*ne*)*škodlivost* ‘(non-)harmfulness’ can be modified by the Patient in the prepositional phrase introduced by *pro* ‘for’ (e.g., (*ne*)*škodlivost pro lidské zdraví* ‘(non-)harmfulness to human health’). Exceptionally, a complement of the N-form takes on a form that is not used with the corresponding A-form at all; for example, while both the A-form *libost* ‘pleasure’ and the N-form *nelibost* ‘displeasure’ can be modified by the Patient in the prepositional phrase introduced by *nad* ‘at’ (e.g., (*ne*)*libost nad jejich počínáním* ‘(dis)pleasure at their actions’), only the N-form *nelibost* ‘displeasure’ can be modified by the Patient in the prepositional phrase introduced by *s* ‘with’, see (1).

(1) neskrývala	nelibost	s	výroky
not-hide- PST-1SG	displeasure-ACC	with	statements-INS
nového	vedení		
new-GEN	management-GEN		
‘(she) did not hide displeasure with the statements of the new management’			

Cases in which N-forms of Czech deadjectival nouns are employed in light verb constructions are relatively rare in our data, e.g., *vyhlásit nezávislost* ‘declare independence’, *vyjádřit nespokojenost* ‘express discontent’, *projevit nelibost* ‘show displeasure’, *vyvolat nevolnost* ‘cause sickness’. N-forms of deadjectival nouns can also be used in idioms such as *pracovní neschopnost* ‘incapacity for work’.

6. Conclusion

Various aspects of word-level negation in Czech nouns and adjectives pose a challenge to lexicographers. In this paper, we have specified some of the issues to be resolved when treating word-level negation in a valency lexicon. Working on the assumption that direct negation of Czech nouns and adjectives should be treated within entries for the corresponding affirmative forms, we have sketched out annotation guidelines and the negation-related attributes applied to the data of the NomVallex lexicon. We have presented preliminary results of the annotation, illustrating these through the case of deadjectival nouns. We believe that this lexicographic treatment will facilitate further research into negation in Czech nouns and adjectives.

References

- Benko, V. (2015). *Araneum Bohemicum Maximum*, verze 15.04. Ústav Českého národního korpusu FF UK, Praha. Available at <https://www.korpus.cz>.
- Cruse, D. A. (1986). *Lexical Semantics*. Cambridge University Press.
- Curiel, M. M. (2015). Negation. In P. O. Müller et al. (Eds.), *Word-Formation. An International Handbook of the Languages of Europe. Volume 2* (pp. 1351–1359). De Gruyter Mouton.
- Dvořák, V. (2016a). Dějové substantivum. In P. Karlík et al. (Eds.), *Nový encyklopedický slovník češtiny* (pp. 309–315). Nakladatelství Lidové noviny.
- Dvořák, V. (2016b). Verbální substantivum. In P. Karlík et al. (Eds.), *Nový encyklopedický slovník češtiny* (pp. 1948–1953). Nakladatelství Lidové noviny.
- Hajič, J. et al. (2003). PDT-VALLEX: Creating a Large-coverage Valency Lexicon for Treebank Annotation. In *Proceedings of The Second Workshop on Treebanks and Linguistic Theories* (pp. 57–68). Vaxjo University Press.
- Haugen, T. A. (2013). Adjectival valency as valency constructions. Evidence from Norwegian. *Constructions and Frames*, 5(1), 35–68. <https://doi.org/10.1075/cf.5.1.02hau>
- Herbst, T. et al. (2004). *A valency dictionary of English: a corpus-based analysis of the complementation patterns of English verbs, nouns, and adjectives*. Walter de Gruyter.
- Joshi, S. (2020). Affixal negation. In V. Déprez & M. T. Espinal (Eds.), *The Oxford Handbook of Negation* (pp. 75–88). Oxford University Press.
- Kettnerová, V., Lopatková, M., Bejček, E., & Barančíková, P. (2018). Enriching VALLEX with Light Verbs: From Theory to Data and Back Again. *The Prague Bulletin of Mathematical Linguistics*, 111, 29–56. <https://doi.org/10.2478/pralin-2018-0006>
- Kolářová, V. (2020). Vztah afirmativní a negované formy adjektiv a substantiv z hlediska jejich valence. *Prace Filologiczne*, 75(1), 293–312. <https://doi.org/10.32798/pf.662>
- Kolářová, V., Kettnerová, V., & Mirovský, J. (2023). Through Derivational Relations to Valency of Non-verbal Predicates in the NomVallex Lexicon. *Jazykovedný časopis*, 74(1), 182–192. <https://doi.org/10.2478/jazcas-2023-0036>
- Kolářová, V., & Vernerová, A. (2022). NomVallex: A Valency Lexicon of Czech Nouns and Adjectives. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)* (pp. 1344–1352). European Language Resources Association.
- Kolářová, V., Vernerová, A., & Klímová, J. (2022). *NomVallex 2.0*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. Available at <http://hdl.handle.net/11234/1-4663> and at <https://ufal.mff.cuni.cz/nomvallex>.
- Kováříková, D. (2011). Gramatická kategorie negace. In F. Čermák (Ed.), *Korpusová lingvistika Praha 2011. 2: Výzkum a výstavba korpusů* (pp. 273–280). Nakladatelství Lidové noviny.
- Kováříková, D., Chlumská, L., & Cvrček, V. (2012). What Belongs in a Dictionary? The Example of Negation in Czech. In R. V. Fjeld & J. M. Torjusen (Eds.), *Proceedings of the 15th EURALEX International Congress* (pp. 822–827). Department of Linguistics and Scandinavian Studies, University of Oslo.
- Křen, M. et al. (2023). *Korpus SYN*, verze 12 ze 29. 12. 2023. Ústav Českého národního korpusu FF UK, Praha. Available at <https://www.korpus.cz>.

Lopatková, M. et al. (2022). VALLEX 4.5. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. Available at <http://hdl.handle.net/11234/1-4756>.

Lotko, E. (1973). *Lexikální negace v současné češtině*. Státní pedagogické nakladatelství.

Pavlovič, J. (2015). Negation in the Slavic and Germanic languages. In P. O. Müller et al. (Eds.), *Word-Formation. An International Handbook of the Languages of Europe. Volume 2* (pp. 1360–1373). De Gruyter Mouton.

Rechzieglová, A. (1995). *On Negation in Czech*. Euroslavica.

Rusínová, Z., & Osolsobě, K. (2016). Prefixace. In P. Karlík et al. (Eds.), *Nový encyklopedický slovník češtiny* (pp. 1394–1395). Nakladatelství Lidové noviny.

Svozilová, N., Prouzová, H., & Jirsová, A. (2005). *Slovník slovesných, substantivních a adjektivních vazeb a spojení*. Academia.

van Son, C., van Miltenburg, E., & Morante, R. (2016). Building a Dictionary of Affixal Negations. In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics (ExProM)* (pp. 49–56). Osaka, Japan.

Šimandl, J. et al. (2016). *Slovník afixů užívaných v češtině*. Karolinum.

Štícha, F. et al. (2013). *Akademická gramatika spisovné češtiny*. Academia.

Zimmer, K. E. (1964). Affixal negation in English and other languages: An investigation of restricted productivity. *Word*, 20, Issue sup. 1, Monograph No. 5.

Acknowledgements

This work was supported by the Czech Science Foundation (project No. 22-20927S). The work described herein has been using data and tools provided by the LINDAT/CLARIAH-CZ Research Infrastructure (<https://lindat.cz>), supported by the Ministry of Education, Youth and Sports of the Czech Republic (project No. LM2023062).

Contact information

Veronika Kolářová

Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Czech Republic
kolarova@ufal.mff.cuni.cz

Jiří Mírovský

Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Czech Republic
mirovsky@ufal.mff.cuni.cz

Kusujiro Miyoshi

JOHN PICKERING'S REFERENCE MATERIALS FOR HIS VOCABULARY (1816)

Transcending the Bounds of Dictionaries

Abstract John Pickering's *Vocabulary* is judged to have been scarcely analyzed in detail, in spite of the fact that it is the first dictionary of Americanisms. (In this respect, the situation seems to be greatly different from the case of Robert Cawdrey's *Table Alphabeticall* (1604), the first English monolingual dictionary in England.) In this situation, it has been found recently that Pickering utilized wide range of dictionaries in the compilation process of the *Vocabulary* and that he may be worthy to be called an originator of comparative lexicography in America. However, there is another historically significant aspect concerning the dictionary. That is, Pickering exercised his ingenuity in providing abundant linguistic information by utilizing quite a few materials other than dictionaries. In this paper, I will deal with the aspect, aiming to further clarify the first stage of American lexicography.

Keywords Pickering; Americanisms; reference materials other than dictionaries; lexicographical practice

1. Introduction

John Pickering, whom I am to deal with in this paper, is the lexicographer who compiled the first dictionary of Americanisms at the beginning of the nineteenth century under the long title *Vocabulary, or Collection of Words and Phrases Which have been Supposed to Be Peculiar to the United States of America* (1816). Read (2002, p. 114), a masterly scholar of Americanisms, acclaims the dictionary as “an important landmark in the study of the English language in America” and Mencken (1982, p. 48), the legendary authority who compiled the historic volume *The American Language* (1919–1948), regards it as the “first really competent treatise on the subject [Americanisms].”

However, as far as I can judge, the content of Pickering's *Vocabulary* has only been slightly analyzed for two centuries since its publication. Actually, the journal *American Speech*, for instance, has carried only five articles that discuss Pickering's *Vocabulary* since its initial issue published in 1925. One of them is Allen Read's *The Collections for Pickering's 'Vocabulary'* (1947). The other four, namely those by Southard (1979), Anderson (1992), Haugen (1992) and Montgomery & Mishoe (1999), merely incidentally touch upon the *Vocabulary* in discussing other topics, respectively.

With regard to books, Julie Andresen's *Linguistics in America, 1769–1924* (1990) is thought to be one of very few ones which are noteworthy. In this book, Andresen seems to make full use of linguistic information provided in the *Vocabulary*, but does not go into the details of the lexicographical practice of its author, Pickering.

Against this backdrop, Miyoshi (2021) has claimed to the effect that Pickering wonderfully utilized a wide range of dictionaries in the *Vocabulary*, and that he can even be seen as an originator of comparative lexicography. In relation to this point, the following is a passage in the entry on **to legislate** of the *Vocabulary*, which I cite to show one aspect of such a situation:

Walker has inserted it [*legislate*] in his dictionary, but (as he remarks) it is “neither in *Johnson* nor *Sheridan*,” nor is it in *Mason*’s Supplement to *Johnson*. It was noticed, however, several years ago in *Entick*’s dictionary, (edition 1795); and, more lately, in an edition of *Sheridan*, “corrected and improved by Salmon;” and also in the octavo edition of *Perry*’s dictionary, published in 1805. Mr. Webster adopts it from *Entick*.

It may safely be said that this passage indicates that my assertion on Pickering’s *Glossary* to be not unreasonable. At the same time, however, there is another fact associated with Pickering’s use of dictionaries. The fact is that Pickering, having profound knowledge on various types of dictionaries, was far from solely based on them in exploring language facts.

For instance, he, referring to descriptions in dictionaries, very frequently goes further than them. The following passage, which is from the entry on **mush**, is one such example; here, he provides information on a dialect which is more detailed than the descriptions in the dictionary he refers to:

“Food of *maiz*, flower and water boiled. (local).” *Webst. Dict.* [Noah Webster’s *Compendious Dictionary of the English Language* (1806)] Used in some of the Southern States, for the same thing that in Northern States is called *Hasty-pudding*.

For one thing, such a fact is thought to have resulted from Pickering’s extensive reading of literature. The passage in the following, which is from the entry on **perk**, will be regarded as supporting the possibility, where he points out Samuel Johnson’s oversight, based on the description in an article in a magazine:

This word [*perk*] is in *Johnson* [*Johnson’s Dictionary of the English Language* (edition unknown)], but is marked “*obsolete*.” It is, however, *provincial* at this day, in England: “*Perk*; lively.” *Specimens of the Essex Dialect, in the Monthly Magazine, for July 1814*, p. 498.

In relation to this example, a passage in the entry on **to notice** may be worth citing, in which Pickering closely compares the descriptions in dictionaries and the language facts:

It [*notice*] is a modern word, and is not in *Johnson*’s dictionary. *Mason* [George Mason’s *Supplement to Johnson’s Dictionary* (1803)] says, it is “a word imported into English conversation from *Ireland*,” but it is now used in *England*, both in conversation and in writing: “This work, which we really thought we had noticed long ago.” *British Critic*, vol. xxxiv. p. 537. “The fourth, which we lately noticed, &c.” vol. xxxv. p. 18. The only English dictionary in which I find it, is *Ash*’s [John Ash’s *New and Complete Dictionary* (1775 edition)]; where it is said to be “*not much used*.” But that work was published forty years ago.

After seeing the examples above, it may safely be said that we are ready to analyze in what ways Pickering used his reference materials other than dictionaries, a subject which is judged to have been almost entirely neglected until today.

For my analysis of the reference materials in this paper, I will adopt a sampling method, treating 113 entries in the *Vocabulary* whose head-words and head-phrases begin with the letters *J, K, L, M, N, O* and *P*, which comprise approximately 18% of all its entries. Within this range, Pickering cites his reference materials, with the exception of dictionaries, 146 times in 70 entries, which account for 61.9% of all entries. And the materials can be divided into seven categories: (1) the English Bible, (2) state papers, (3) private letters, (4) the record of lectures and sermons, (5) newspapers, (6) periodicals and (7) other books.

On the premises as above, I will divide my analysis into three sections, dealing with Pickering's way of using the reference materials of categories (1) to (4) at first, determining them as "type 1", that of the categories (5) and (6) as "type 2" after that and, thirdly, the category (7) as "type 3", which is due to the difference of his way of treating the reference materials. By adopting such an inductive method (or "seeing trees"), I aim to clarify Pickering's intention to have referred to a wide range of literature in the concluding section (or "discerning the entire picture of a forest").¹

2. Use of Reference Materials Type 1: The English Bible, State Papers, Private Letters and the Record of Lectures and Sermon

If I am to show the specific of each of the materials, indicating how often Pickering refers to it, the situation is as the following table shows.

Table 1: Pickering's reference materials type 1

Categories (Frequency)	Specifics
The English Bible (3 times in all in 1 entry)	-----
State Papers (8 times in all in 8 entries): 8 such papers	<i>Debates in Parliament</i> (1 time), <i>Debates on the Judiciary</i> (1 time), <i>Duke of York's General Orders</i> (1 time), <i>List of Candidates at Election in Rhode Island</i> (1 time), <i>Massachusetts Colony Laws</i> (1 time), <i>Massachusetts Provincial Statutes</i> (1 time), <i>Massachusetts Statutes</i> (1 time), and <i>Report of the Secretary of State on Commercial Restrictions</i> (1 time).
Private Letters (6 times in all in 4 entries): those of 5 persons	Edmund Burke (1 time), Lord Chesterfield's (1 time), Thomas Jefferson's (1 time), Junius's (2 times in all in 2 entries), and John Adams's (1 time).
The Record of Lectures and Sermons (4 times in all in 3 entries): those of 4 persons	Benjamin Waterhouse's "Lecture on the evil tendency of the use of Tobacco", George Hill's "Institutes of theology" (1 time), George Whitfield's sermons (1 time), and Steven MacGill's "Considerations addressed to a young clergyman" (1 time).

¹ Apart from the above, at the end of this introductory section, it should be clarified that this paper is to follow up my three previous papers on Pickering (Miyoshi, 2016; 2021a; 2021b). As to the general pictures of Pickering and his *Vocabulary*, I hope these papers would be referred to.

At this first stage of my analysis, it will already be inferred that Pickering’s *Vocabulary* is not simply a collection of Americanisms, he referring to *Duke of York’s General Orders* and Lord Chesterfield’s letter which are materials published outside of America; this is because Pickering aimed to reveal the facts of Americanisms by using materials published in Britain, as well, which will be more noticeable in the following sections.

The reason I lump together the four categories of reference materials is that Pickering used them in one single way, that is, he, in the *Vocabulary*, always illustrates the actual use of words and phrases with them; this is different from the case of newspapers, periodicals and other books which I will deal with later. To cite one instance from each of the four types of materials, it is as the following; here, we can see that Pickering usually gave full attention to making clear the sources of his materials:

1. The English Bible: [from the entry on **plead or pled for pleaded**] “‘There I will plead with you [...] like as I *pleaded* with your fathers [...]’ *Ezek.* xx. 35, 36 [...]”
2. State Papers: [from the entry on **kental**] “‘Our whale-oils pay six livres the *kental*.’ *Report of the Secretary of State* (Mr. Jefferson) *on Commercial Restrictions &c.* Dec. 16, 1793 [...]”
3. Private Letters: [from the entry on **occlusion**] “‘The *occlusion* of the port of New Orleans [...]’ *Letter of President Jefferson to Gov. Garrard, Dec.*”

3. Use of Reference Materials Type 2: Newspapers and Periodicals

As to the specific of the materials type 2, it is as shown in the following table.

Table 2: Pickering’s reference materials type 2

Categories (Frequency)	Specifics
Newspapers (4 times in all in 3 entries): 2 newspapers	Connecticut newspaper (title not mentioned) (1 time), <i>New York Evening Post</i> (2 times in all in 2 entries) and <i>Salem Gazette</i> (1 time).
Periodicals (58 times in all in 38 entries): 15 periodicals s	<i>Annual Review</i> (2 times in all in 2 entries), <i>British Critic</i> (22 times in all in 14 entries), <i>Cambridge Literary Miscellany</i> (1 time), <i>Christian Observer</i> (3 times in all in 3 entries), <i>Collections of the Massachusetts Historical Society</i> (1 time), <i>Edinburgh Reviewers</i> (4 times in all in 4 entries), <i>European Magazine</i> (1 time), <i>London Monthly Magazine</i> (1 time), <i>Missionary Magazine</i> (1 time), <i>Monthly Anthology</i> (9 times in all in 9 entries), <i>Monthly Magazine</i> (1 time), <i>Port Folio</i> (2 times in all in 2 entries), <i>Quarterly Review</i> (7 times in all in 7 entries), <i>Spectator</i> (2 times in all in 2 entries) and <i>Tatler</i> (1 time).

Here we see Pickering refer to various kinds of periodicals, in addition to a few newspapers, such as those on religion, literature and politics; he also refers to the *Spectator* and the *Tatler*, renowned periodicals in the history of English literature. Besides, it is to be noted that the *British Critic*, a conservative and high church review journal established in 1793, is by far the most frequently cited periodical, an American one the *Monthly Anthology*, a miscellaneous magazine which was first issued in 1804,

being cited next to it in terms of frequency; he also refers to *The Edinburgh Review*, a periodical in Scotland.

With regard to the materials of this type, Pickering used them for two purposes. The one is to illustrate the use of words and phrases, as we have seen in the case of the type 1, and the other to show opinions and views on it by authorities and commentators, which he sometimes supports and sometimes criticizes. Specifically, within the range of our current scope, Pickering refers to the materials 22 times for the former purpose and 40 times for the latter.

Here, to cite a few examples, focusing on the case where the materials are used to show authorities' and commentators' opinions and views, they are as the followings. As to the use of newspapers, in the entry on **over for under** of the *Vocabulary* reads like this:

In these expressions: "He wrote *over* the signature of Junius. He published some papers *over* his own signature." / A few of our writers still countenance this unwarrantable innovation [...]. The use of the word *under*, is as well established as any English idiom. [...]. / Mr. Coleman, the able editor of the New York Evening Post, has repeatedly exposed this "piece of affectation," (as he justly calls it,) and produced the following authorities [...].

Then, Pickering places a footnote for the passage as the following:

See the *New York Evening Post* of March 15, and Nov. 22, 1803. A writer in another newspaper, who adopts the signature of *The Good Old Way*, ironically closes his remarks upon this "awkward and absurd term," as he styles it, by thus employing it --- "Given *over* my hand and seal," &c. *Salem Gazette*, Apr. 2, 1813.

As for the use of periodicals, Pickering, for instance, uses them to clarify the origins of Americanisms from the perspective of dialectology, as seen in the two examples below:

From the entry on **kelter or kilter**:

Good condition, order. *Ex.* This cart, or plough, is out of *kilter*. / This is very common among the farmers of New England. It is also one of *provincial* words of Great Britain [...] in the *Monthly Magazine* (Mar. 1815) it is given among specimens of the *Essex* Dialect.

From the entry on **old for stale**:

[...] in this expression, *old bread*. *New England*. From the following extract, this seems also to be a *Scotticism*: "The *Scotticism* *old bread* seems no way inferior to the *Anglicism* *stale bread*." *London Monthly Magazine*, Apr. 1800, p. 239 ago.

In this regard, Read (2002) has once commended Pickering as an “astute philologist”, with reference to his paper in 1828, remarking the following:

Their (of speakers from Scotland and Ireland) influence on American speech was suggested in 1828 by the astute philologist John Pickering: “We have already had in this country great numbers of Scottish and Irish teachers; and they, together with Scottish Dictionaries and Spelling Books, have had no inconsiderable influence in the Middle States, upon our pronunciation and language” (*American Quarterly Review*, Sept. 1828, 208). (Read, 2002, pp. 88–89)

Back to Pickering’s reference to periodicals, he was not necessarily obedient to descriptions in them, as the following example indicates:

From the entry on **pending**:

In the review of *Marshall’s Life of Washington* in *the Monthly Anthology* (vol. v. p. 438.) this word is criticised as *obsolete*. But it is certainly used by the English in *legal* and *parliamentary* language; and it is, accordingly, to be found in the Reviews of works on those subjects: “At the period when the American treaty with this country was *pending*,” &c. *Briti. Crit.* vol. vi. p. 594

4. Use of Reference Materials Type 3: Other Books

The following table shows the specifics of materials type 3, other books; differently from the cases of other types of materials, Pickering basically does not indicate the year of publication concerning them, the reason for which may be ascribed to the possibility that he thought his readers could find a book of this type by author and title.

Table 3: Pickering's reference materials type 3

Other Books (63 times in all in 44 entries): 28 books	
American Academy of Arts and Sciences, <i>Memoirs of the American Academy of Arts and Sciences</i> (1 time).	-----, <i>Rural Economy of Norfolk</i> (2 times in all in 2 entries).
Barnes, Thomas, <i>Parliamentary Portraits</i> (1 time).	-----, <i>Rural Economy of the Midland Counties</i> (1 time).
Burke, Edmund, <i>Reflections on the Revolution in France</i> (American edition) (4 times in all in 2 entries).	-----, <i>Rural Economy of Yorkshire</i> (American edition) (1 time).
Campbell, George, <i>The Philosophy of Rhetoric</i> (Boston edition) (2 times in all in 2 entries).	-----, <i>Rural Economy of Yorkshire</i> (British edition) (1 time).
Edgeworth, Maria, <i>Castle Rackrent</i> (1 time).	Pegge, Samuel, <i>Anecdotes of the English Language</i> (1 time).
Franklin, Benjamin, <i>Franklin's Essays</i> (3 times in all in 3 entries).	Priestley, Joseph, his grammar (title not specified) (1 time).
Hargrave, Francis (ed.) <i>A Collection of Tracts Relative to the Law of England</i> (1 time).	Ramsay, David, <i>History of the American Revolution</i> (American edition) (2 times in all in 2 entries).
Imlay, Gilbert, <i>Topographical Description of the Western Territory of North America</i> (1 time).	-----, <i>History of the American Revolution</i> (British edition) (2 times in all in 2 entries).
Kendal, Augustus, <i>Travels through the Northern Parts of the United States</i> (11 times in all in 11 entries).	-----, <i>History of the Revolution of South Carolina</i> (British edition) (1 time).
Lambert, John, <i>Travels through Lower Canada, and the United States of America</i> (1 time).	Tooke, Horne, <i>Diversions of Purley</i> (American edition) (1 time).
Lowth, Robert, his grammar (title not specified) (2 times in all in 2 entries).	Webster, Noah, <i>Dissertations on the English Language</i> (2 times in all in 2 entries).
Marshall, John, <i>The Life of Washington</i> (American edition) (12 times in all in 10 entries).	Williams, Roger, <i>A Key into the Language of America</i> (1 time).
-----, <i>The Life of Washington</i> (British edition) (1 time).	Witherspoon, John, <i>Druid</i> (5 times in all in 4 entries).
Marshall, William, <i>Rural Economy of Gloucestershire</i> (1 time).	Woodfall, Henry (ed.), <i>Miscellaneous Letters of Junius</i> (Woodfall's "Advertisement" prefixed to its body) (1 time).

Within this range, Pickering is seen to refer to the books 24 times to illustrate the actual use of words and phrases and 39 times to cite opinions and views on it by authorities and commentators.

When browsing through Pickering's *Vocabulary* with reference to the books shown in the table above, we can find several notable facts. For instance, he uses authoritative reference materials on Americanisms as John Witherspoon's *Druid*, as well as refers to descriptions about the relevant language information in such books as Gilbert Imlay's *Topographical Description of the Western Territory of North America* and Augustus Kendal's *Travels through the Northern Parts of the United States*, giving the impression that the *Vocabulary* is a collection of essential information on the language before Pickering.

Besides, he has collected extensive dialectal data in Britain to explore the background of Americanisms, using books such as Thomas Barnes's *Parliamentary portraits*, Francis Hargrave's *Collection of Tracts Relative to the Law of England* and William Marshall's four titles on Gloucestershire, Norfolk, Yorkshire and midland counties in England. Furthermore, Pickering even uses Roger Williams's *Key into the Language of America*, a seventeenth-century description of the Native American language Narragansett.²

However, if I take up the characteristics peculiar to Pickering's use of reference materials type 3, other books, which are not seen in the case of other types of materials, he uses the American and British versions of the same title and refers to classic books on the English language. The situations of the two points are, respectively, as the followings.

As to the former case, Pickering's use of the American and British versions of the same title, in the entry on **to originate**, for instance, he, after commenting on the objection by general British people about the use of the word *originate* as the transitive verb, cites two phrases from John Marshall's work, the one from the American edition and the other from the London edition, as shown below:

[...] in *Marshall's Life of Washington* (vol. i. p. 33) where the *American* edition has this expression --- "Bartholomew Gosnold, who *originated* the expedition" --- the *London* edition has --- "who had *planned* the expedition."

To cite another similar example, in the entry for the word **occlusion**, which is defined as 'a shutting up, closing', Pickering refers to a sentence from David Ramsay's work, as:

It [*occlusion*] was used [...] in Dr. Ramsay's *History of the American Revolution*, (published in 1789) vol. i. p. 103: "He had also hoped, that the prospect of advantage to the town of Salem, from its being made the seat of the custom-house, and from the *occlusion* of the port of Boston, would detach them from the interest of the latter," &c.

Then, under this citation, he says:

In the *London* edition of the work, this word [*occlusion*], being doubtless new to the English editor, was probably supposed to be an error of the press in the *American* copy, and it is accordingly changed into a word resembling it in sound, and would occupy the same space in the page, the word *ex-clusion*.

Concerning Pickering's reference to classic books on the English language, the entry on **mean for means**, one of the longest entries in his *Vocabulary*, being comprised of 71 lines in 3 pages, begins with these words: "Many American writers, following the Scottish models, make use of *mean* instead of *means*. *Ex.* 'It was the best *mean* of bringing the negotiation to a happy issue.' *Marsh. Life of Washington*, vol. v. p.

² In association with this point, Witherspoon's *Druid* which I have cited is originally a periodical, but here I treat it as one of "other books" for the reason that it exclusively provides the analysis of Americanisms, being different from other periodicals which I have treated as belonging to type 2 in Section 3.

546.” In this entry, for the purpose of examining the correctness of the use of *mean* as a synonym of *measure* and *way*, Pickering goes into detail on various opinions and views of historically famous authorities on the language, minutely indicating the respective sources of their relevant books, apart from the year of publication, as the following:

Robert Lowth's “*Grammar, ch. on Sentences, note 1*”, Joseph Priestley's “*Grammar, Notes, sect. 1*”, George Campbell's “*Philos. of Rhet [Philosophy of Rhetoric]. B. ii. ch. 2. sect. 2. p. 216, Boston edit.*” and Horn Tooke's “*Diversion of Purley, part 2.47. Amer. edit.*”

The situation similar to this is also seen in the entries on **lit or light, plenty for plentiful, etc.**

5. Conclusion

To summarize my analysis in this paper, Pickering used his reference materials for three purposes in his *Vocabulary*: (1) for illustrating the use of words and phrases; (2) for showing opinions and views on words and phrases by authorities and commentators; and (3) for indicating the differences of American and British ways of using words and phrases, providing abundant dialectal information.

For these purposes, he, within the range of approximate 18% of entries in the *Vocabulary*, has greatly made use of the English Bible, 8 state papers, the private letters of 5 historical persons, the records of lectures and sermons by 4 persons, 2 newspapers, 15 periodicals and 28 books, and this has all been done, filling up the deficiency of dictionaries which had been published before him.

After having known the amount of background information on Americanisms which Pickering revealed in the 1810's, using such reference materials, Read's words of acclaim that the *Vocabulary* is “an important landmark in the study of the English language in America” (Read, 1947, p. 271), which I cited at the beginning of this paper, seem to be still insufficient.

Here, we may remember Samuel Johnson's famous saying on many people's lips: “Dictionaries are like watches; the worst is better than none, and the best cannot be expected to go quite true”. Pickering, who has crossed the bounds of dictionaries, may be one of very few persons that fully recognized the meaning of this saying.

References

- Andresen, J. T. (1990). *Linguistics in America, 1769–1924: A Critical History*. Routledge
- Anderson, E. R. (1992). Americanisms in Mary Botham Howitt's *Our Cousins in Ohio* (1849). *American Speech*, 67(3), 326–330.

Haugen, E. (1992). A language survey that failed: Seip and Selmer. *American Speech*, 67(3), 330–336.

Mencken, H. L. (1982). *The American Language: An Inquiry into the Development of English in the United States* (the fourth edition and the two supplements, abridged, with annotations and new material, by McDavid, R. I. Jr.) (one-volume abridged edition). Alfred A. Knopf. Inc.

Miyoshi, K. (2016). Reconsidering the initial development of the dictionaries of Americanisms: Pickering's legacy in Bartlett's *Dictionary* (1849). In T. Margalitadze, & G. Maladze (Eds.), *Proceedings of the XVII EURALEX International Congress: Lexicography and Linguistic Diversity* (pp. 597–607). Ivane Javakhishvili Tbilisi State University.

Miyoshi, K. (2021a). John Pickering's *Vocabulary* (1816) reconsidered: America's earliest philological exploration of lexicography. In Z. Gavrilidou et al. (Eds.), *Lexicography for Inclusion: Proceedings of the XIX EURALEX Congress* (pp. 315–320). Democritus University of Thrace.

Miyoshi, K. (2021b). Pickering's influence on Craigie and Hulbert's *Dictionary of American English* (1936–1944). *Lexicographica: International Annual for Lexicography*, 1(1), 249–259.

Montgomery, M., & Mishoe, M. (1999). "He bes took up with a yankee girl and moved up there to New York": The verb *bes* in the Carolinas and its history. *American Speech*, 74(3), 240–281.

Pickering, J. (1997). *A Vocabulary or Collection of Words and Phrases Which Have Been Supposed to Be Peculiar to the United States of America* (facsimile reprint of the original edition published in 1816). Routledge/Thoemmes Press.

Read, A. W. (1947). The collections for Pickering's 'Vocabulary'. *American Speech*, 22(4), 271–286.

Read, A. W. (2002). *Milestones in the History of English in America* (Publication of the American Dialect Society, No. 86) (Annual Supplement to *American Speech*). Duke University Press.

Southard, B. (1979). Noah Webster: America's forgotten linguist. *American Speech*, 54(1), 12–22.

Acknowledgements

In writing this paper, I was given valuable suggestions from Professor Emeritus John Patrick Considine (University of Alberta). However, remaining errors, if any, are my responsibility.

Contact information

Kusujiro Miyoshi
Soka Women's College
KusujiroMiyoshi@gmail.com

Pär Nilsson

REPORT ON THE REVISION OF THE SWEDISH ACADEMY DICTIONARY – AND THE SEARCH FOR “OLD NEOLOGISMS”

Abstract In 2023, the first edition of the diachronic-contemporary *Swedish Academy Dictionary* (SAOB) was finished after more than 130 years of work. The dictionary contains over 500,000 entries and describes the Swedish language from the 1520s to the present day. However, a consequence of the extended time of publication is that a significant part of the content, in particular in the first volumes, was out of date even by the time of completion of the full edition. Headwords had become obsolete, the meta-language was outdated in several regards, and a lot of modern words are lacking in the lemma list of the dictionary. To address this, a revision project was initiated in the beginning of 2024. This article reports on the revision in more detail, with emphasis on the supplement of new entries.

Keywords Swedish Academy Dictionary; historical dictionaries; dictionary revision; neologisms

1. Introduction

Up to 2023, the *Swedish Academy Dictionary* (SAOB) was sometimes said to be the longest ongoing non-finished dictionary project in the world. Correct or not, the project was initiated as early as in the 18th century, but it did not get its lexicographical breakthrough until the last decades of the 1800s when a permanent editorial board was installed in Lund. Under this new arrangement, the first volume of the dictionary was published in 1898 after 10 years of preparations (cf. Malmgren, 2011). Since then a new volume has been published on average approximately every three or four years. Of course, there were some worries about the pace of publication throughout the years, and more than once it has been questioned if the dictionary were ever to be finished (see e.g., Lundblad (2003, p. 100) who refers to discussions on this topic in the first decades of the 1900s). However, since 2023 the first edition of SAOB is officially completed. By the end of that year the 39th volume, containing words beginning with the diacritic letters Ä and Ö, was printed and launched on the dictionary website <https://www.saob.se/>. But since the work has progressed only alphabetically forward, with no revision at all for older parts of the dictionary, a lot of entries, in particular in the first volumes, are out of date already, either because the words themselves have become obsolete or else the actual phenomenon that the notions denote. As the result of the long publication time many words and meanings, coined after the publication of the current alphabetic span, are also missing. In addition, some of the meta-linguistic terms and expressions used in the dictionary have become outmoded. Hence, there is room and reason for revision to be done. Therefore, a revision project was started at the beginning of 2024 with the goal of achieving a more balanced dictionary, without obvious gaps in the list of lemmas and with a modernized meta-language.

The aim of this paper is to describe the revision project and its content further, with the main focus on the addition of new entries and new headwords, i.e., the part of the dictionary work that has progressed the farthest at the time of writing. In order to grasp the nuances of this description more easily, a substantial background description of SAOB is also given. The paper is organized as follows: Section 2 describes the SAOB dictionary project, from a historical perspective as well as in terms of content, macro- and microstructural organisation etc. In Section 3, the general principles and different parts of the revision project are outlined. Section 4 deals with one of these parts in more detail, namely the neologisms (in SAOB terms), and in Section 5, lastly, some concluding remarks are made.¹

2. The Dictionary – a General Background

When the Swedish king Gustav III founded the Swedish Academy in 1786, it was declared in the Statutes (§ 23) that “it is [...] incumbent on the Academy to compile a Swedish dictionary”.² The following year the Academy members themselves initiated the dictionary work simply by distributing the letters of the alphabet within the group, after the model of the French Academy (Svensén, 2015, p. 87). For the Swedish Academy, however, this approach never led to a finished dictionary. As already indicated, the method of working did not find its form until an external editorial board with recognized lexicographical competence was installed in the 1880s. As a scientific leader of the project, Sweden’s most prominent lexicographer at the time, K. F. Söderwall, by then just in the process of finishing his solely historical dictionary of the Swedish medieval language (OSM), was chosen. Given the scope of his completed dictionary, it was decided that the period of description for SAOB was to begin where OSM ended, namely by the 1520s. (SAOB’s oldest source is the first volume of Gustav I’s registry from the year 1521, which thus marks the boundary between old and modern Swedish in the dictionary.)

The first task was to draw up a detailed plan of how the work was to be carried out and what source material would be selected and excerpted. This was followed by several years of collection before sufficient material was available in the archives to begin the first part of the first volume of the dictionary, which was eventually published in 1893 (the first full volume was completed 1898). A great deal of the original plan in terms of layout and content was then valid until the completion of the entire dictionary work. Svensén (2015, p. 100) points out, for example, that even current principles for the selection of compounds and derivatives or for technical terms were already specified in this plan.

2.1 Type, Target Groups and Organization of Entries

SAOB has a historical orientation. But unlike, for example, the previously mentioned Söderwall’s OSM, SAOB also has a contemporary perspective (according to Svensén’s (2009) typology, SAOB can be categorized as a diachronic contemporary linguistic

¹ Thanks to two anonymous reviewers for providing detailed and most helpful comments and suggestions.

² <https://sa-admin.lb.se/assets/3c46295e-685c-4bf1-a39d-61f1d8cd6f85.pdf>

dictionary; cf. Nilsson (2023, p. 25f) for a more complete typology of SAOB based on Svensen's terms). However, the contemporary perspective is relative – the contemporaneity in relation to later readers has, for obvious reasons, changed over the years – but the aim has always been to describe the words (their forms and sense development) from the 1520s to the current date.

The historical principle entails that the chronologically oldest meaning (globally speaking, not necessarily the first one encountered in Swedish) is presented first, that every lexical sense and micro-sense is supported by an authentic quotation, and that each meaning section in the dictionary is sampled by the oldest quote representing the current section. The same principle applies to subordinate meaning sections, whose first quote in order is the oldest found in the editors' collections.

Not only the perspective of time (diachronic-contemporary), but also the target group of users rests on two legs. The entries contain a lot of information and fairly advanced descriptions, but are still not aimed only at advanced users. In the preliminary preface, which was appended to the first half volume issued of the dictionary, it is stated: "that the dictionary is intended not merely for scientists, but for the entire knowledge-seeking Swedish public" (available: <https://www.saob.se/om>; my translation). This dual focus can still be said to apply, since the following is stated in the current description of the dictionary on the website [saob.se](https://www.saob.se): "Anyone interested in words and expressions will benefit from the dictionary, which is also an important aid in scientific research and a rich source for researchers or anyone who may want help when reading, in particular, older texts" (again in my translation).

The organization of entries, their macro- and microstructure, is clearly marked by its time of origin, both in terms of the printed medium (with a lot of space-saving principles) and the historical perspective, which applied to most, or all, linguistic work at the time the SAOB project in its current form was initiated (cf. Teleman, 2015). A complete account of lexical semantics and morphology, thus, more or less by necessity entailed a chronological account.

One important impact of these organization principles in SAOB, is that secondary lemmas, in terms of word-formation logic, are presented in subordinate divisions (subentries) under their respective primary headwords. Hence, the entry of simplex VAX ('wax') contains subentries like the compound VAX-DOCKA ('wax doll') and the derivative VAXIG ('waxy'). The arrangement has space-saving reasons (the full account of forms and the comprehensive etymological derivation are given in the headword entry only and relevant parts are not repeated in the subentries), but is also grounded on the logic-chronological idea that a simplex word is more original than its compounds and derivatives (see Eaker (2006) for a closer description of SAOB's order of presentation). This kind of macrostructure is sometimes labelled *nest-alphabetical* (cf. Svensén, 2009). The term refers to the "nests" that are formed when morphologically related words (headwords, compounds, derivations, particle verbs, etc.) are brought together. The subordinated lemmas (the nest-lemmas) can in turn include their own compounds and derivatives, which gives the macrostructure a hierarchical character.

Also at the microstructural level, hierarchy is applied in the SAOB entries. A main sense (marked with an Arabic numeral) can have a subordinated sense (small italics) which in turn can have a subordinated sense (small Greek italics) that can have a subordinated sense (small Greek italics + prime symbol (')). The very first part of each entry, the head of the entry, consists of form and etymological descriptions. This section is often very dense in information (and full of space-saving abbreviations), particularly in headword entries. As mentioned earlier, the completed dictionary is published in print, but the corresponding content of the SAOB is also available in digital form. In Figure 1, the microstructure of a headword entry in the digital version is illustrated:



Fig. 1: Demonstration of SAOB's microstructure

2.2 Completion of the First Edition

As mentioned above, the work on the dictionary proceeded with a relatively similar result in terms of the entries' layout since the end of the 19th century and onwards. But in terms of detail and precision, and thus the speed with which the endpoint was approached, the work has changed character over the years. According to the original plan, the editorial board expected that their work would be completed – and that the first subscribers who signed up would see a complete work on their bookshelves – within a period of 10 – 12 years (Svensén, 2015, p. 99). However, the editors did not realize what an extensive task they had undertaken. In fact, this estimated time would have to be multiplied by 10; in total, the project lasted more than 130 years.

Several times the SAOB project has been threatened with closure, and the dictionary work has subsequently had to be reorganized more than once. Larsson (2014) identifies four periods in the history of SAOB that represent different levels of scope and ambition. In the fourth and final period (2000 – 2023), an approach was applied (as a reaction to the period before) that resulted in relatively concise and less detailed descriptions and definitions.

The completed dictionary comprises 39 volumes (including a half-volume with a list of the roughly 24,000 different sources cited in the dictionary) which include more than 500,000 lemmas distributed over more than 33,000 pages. As mentioned above, the dictionary is also available in an identical digital edition, which was first introduced in the 1990s but was re-digitized in the 2000s and is available, since 2016, on the website *saob.se*. Since 2017, the dictionary is also available on the dictionary portal *svenska.se* with its sister dictionaries SO ('The Contemporary Dictionary of the Swedish Academy') and SAOL ('The Swedish Academy Glossary'). The completion of the first volume of SAOB was officially announced in December 2023 at the Swedish Academy's Grand Ceremony in the presence of HRH King Carl XVI Gustaf.

3. Revision

Several years before the completion, it was clear that the dictionary would be in need of revision. In the last years of work, a revision project was therefore planned in parallel with the completion. Part of this preparatory work consisted of what was internally called "a revision tour". During this tour, universities and other research institutes were visited throughout Sweden (with stops in Lund, Växjö, Gothenburg, Stockholm, Uppsala, and Umeå) and in Finland (Helsinki). During the regular linguistic research seminars, various thematic aspects (one per institution) of an intended revision were discussed, and in the invitation to the seminar participants, the historically unique opportunity to influence the formation of SAOB's revised edition was highlighted. In addition, various aspects of the revision were presented and discussed at conferences and symposia in the Nordic countries, and a special half-day seminar with this theme, and with invited speakers, was arranged in connection with the conference *Svenska språkets historia* ('The history of the Swedish language') at Lund University in autumn 2021. During these seminars and presentations, topics such as grammar in the dictionary, material and sources, way of presentation, improvements of the microstructure, technical language and Finland-Swedish words and expressions (approx. 5 % of the population in Finland has Swedish as their first language) were discussed.

Many important points of view were collected, not least in terms of readability, gender and pronunciation designations. All in all, the editors got several clear and concrete ideas and more general input to serve as a point of departure for further work. However, to what extent these acquired insights and points of view could be applied in practice, was, of course, dependent on the financial framework.

3.1 A Brief International Outlook

There are close lexicographic equivalents to SAOB in some other European languages. In this context, *Deutsches Wörterbuch von Jacob und Wilhelm Grimm* (DWB), *Woordenboek der Nederlandsche Taal* (WNT), the *Oxford English dictionary* (OED) and *Ordbog over det Danske Sprog* (ODS) can be mentioned in particular. The first three of these were initiated before the SAOB project started. Hast (1983) claims that WNT and OED in particular served as role models for the SAOB editors at an early stage.

At the time of the completion of SAOB, the first editions of these sister dictionaries had long since been published. In several cases, supplements have also been added or other revision work has begun; see e.g., Scheider (2001) for a description of the revision of DWB, Kinable (2006) for revision aspects related to the handling of quotations in WNT, and Trap-Jensen (2019) of the retrodigitized ODS. SAOB's situation is thus, of course, not unique, even if the number of similar projects is limited. As in the initial stage of the project, the editors now have the opportunity to turn their eyes to these dictionary projects to learn lessons and examine how similar problems have been dealt with. Not least relevant for SAOB is, again, the work carried out by the editors of OED, both in view of the very similar micro- and macrostructures of the two dictionaries, and because editing systems from the same software company are used (namely IDM, cf. Section 3.3 below and see Nilsson (2024b) for a somewhat more detailed description). With regard to OED, cf. also Simpson (2020) and Elliott & Williams (2006).

3.2 Financing

Returning to the Statutes mentioned in Section 2, the Swedish Academy could be said to have fulfilled its task with the completion of the first edition of SAOB, and therefore they had no formal obligation to continue the publication. In addition, the Academy also publishes two other dictionaries (SO and SAOL, mentioned above, the latter of which was originally based on the same source material as SAOB). From the early 2000s, there was an economic agreement between the Swedish Academy and the Swedish state, which to a great extent financed the publication of SAOB. But this agreement ended with the completion of the first edition of the dictionary, and the Swedish Academy announced in 2021 that it could no longer afford publication. This was followed by a period of uncertainty regarding the future of SAOB, before the course turned again in 2022, when a collaboration was initiated between several different foundations which would enable the financing of a revision project corresponding to approx. 7 years of work starting in 2024. The Swedish Academy is still one of the financiers, but now together with the Helge Ax:son Johnson Foundation, the Society of Swedish Literature in Finland, The Royal Swedish Academy of Letters, History and Antiquities, and the Bank of Sweden Tercentenary Foundation.

3.3 A Three-Legged Revision Approach

Based on the thorough preparatory work, the editors had a relatively clear image of what is in need of update in the dictionary. But given the financial framework, the relevant questions to pose are: What can be done in this relatively limited period of time? How much is it possible to revise? The task is thus to accomplish as much as possible with as high standard as possible during the given period. This means, obviously, that not everything which is desirable to change can be changed during the project period. On the other hand, further improvements can possibly be made at a later stage, if the project is extended beyond these 7 years.

Early in the process, three distinct areas, or three different ways of approaching the revision, were identified. The main principles behind these approaches are summarized below.

Firstly, there are certain elements, or aspects, recurring in all or many entries that can be managed in more or less the same way at the same time. This group of changes, and the assignments related to it, is labelled *aspect revision*, and it includes heterogeneous types of aspects to deal with. A prerequisite for this task is the implementation of a new software (Digital Publishing System (DPS), provided by the French company IDM; cf. e.g., Arts, 2014; Kocjančič et al., 2006). The content of the previous digital edition has been transferred to this system and re-tagged. The result is a completely digital database, instead of the previously only digitized version of the printed text.

Optimally, the various aspects can be handled automatically by replacing one kind of element with another. But in practice, this is rarely the case, as most of the tasks require some form of manual effort. Most straight forward are tasks such as untangling abbreviations that are difficult to interpret, and replacing obsolescent spellings (which do not coincide with correct forms of other words) among lemmas and meta-language.

More demanding is the work of replacing the old gender designations from the first edition with modern equivalents (both the gender system itself and the meta-terminology have changed during the time of publication of SAOB). The issue here is that the detailed descriptions (often including several different fonts) are difficult for the software to decode, and more advanced descriptions have to be translated manually. However, many gender designations are more concise and can be translated in a satisfactory way automatically. The editors' approach here is to prioritize the unproblematic instances and make the entries in which they are found as up-to-date as possible in this respect, and then deal with the more complicated cases later, even if this implies that SAOB will, for a period of time, contain entries with two different systems of denoting gender (one based on the older system's combinations of grammatical and sexual gender, and one based on grammatical gender alone; see Nilsson (2024b) for an account of SAOB's way of dealing with gender in the first edition and in the revised version).

More or less on the same level of difficulty is the task of rearranging the content of the definition text based on the hierarchy that already exists in the entries; i.e., SAOB has a subtle way of separating different sections within a definition using punctuation marks and certain keywords. A semicolon plus the word "särskilt" ('especially') indicates that what follows constitutes a subordinated special use of what was previously said, while a semicolon plus "älv." ('also') refers to a more independent secondary sense. By automatically dividing the definition text with a new line for each semicolon and also a short indent for descriptions beginning with "särskilt", the hierarchy can be highlighted and clarified, and at the same time the text becomes less compact. This can be (and has in fact been) done automatically. But a complicating circumstance for the machine interpretation is that the key words "särskilt"/ "älv." can be modified and combined in various ways (e.g., *formerly also*, *sometimes also*, *except*

for older conditions now especially of ..., etc.) and semicolons can (with or without difference in meaning) be replaced by commas or parentheses. This aspect, too, can thus be classified as semi-automatic.

Even more demands on manual effort are required by the task that has to do with the description of semantic relations (see Nilsson, 2024a). Labels such as *transferred use* have been consistently used in the dictionary, but partly in different ways, which will now be adjusted. Such labels can be easily identified in the editing system, but replacing an inappropriate label requires a certain amount of manual work and individual assessment (based on certain criteria).

Secondly, and as a means of compensating for the deficiencies of the aspect revision, *manual revision of the individual entries* is needed. This part of the revision project is still in development. At the time of writing, a pilot study is underway. Manual efforts can of course be made in different ways, and possibly several different methods will be applied later. But since many entries contain a lot of old-fashioned or obsolescent elements – both linguistically (especially in the meta-language; cf. e.g., how plural verb conjugation occurs in SAOB entries despite the fact that the usage disappeared in Swedish in the middle of the 20th century) and regarding the phenomena described (in the entry BÄVER ('beaver') this species is described as extinct in Sweden, which is no longer the case) – it is urgent not to concentrate on a few selected single entries, but to review as much of the dictionary's content as possible. Hence the strategy at present is to focus on the low-hanging fruit, on as wide a range as possible, both in terms of the number of entries and the different types of updates within each entry.

Naturally, the older entries are in most need of revision, and therefore entries with initial letters from A–R (written between the 1890s and 1960s) are selected in this respect. The aim is to go through all entries but with a light hand. Each entry is revised in two levels and by two different editors – one editor reads the whole text and changes outdated wording and inaccuracies in the entries and adds new senses (if there is an obvious need). More complicated cases are skipped for the time being, which means that restructuring of large entries is not done at this stage. The task of the other editor is to supplement entries with modern quotations. This is a relatively uncomplicated task, but one that affects the reader's impression quite a lot. Often it is the lack of examples in a modern context that makes words described in the 19th century perceived as outdated, and just the addition of citations from the 21st century can thus significantly modernize an entry. At the moment, the editors work their way through the alphabet in pairs, with the equivalent of 100 columns in the printed dictionary at a time, which seems a manageable amount. The two editors then review each other's work. It should be added that in the manual revision the chance to adjust the outcome of the aspect revision is also given, which makes this task a kind of control station.

Thirdly, SAOB's long publication time has implied that many modern words are missing, especially in the first half of the dictionary. An area in obvious need of revision is, thus, the *addition of new entries*, the part of the project that is internally called *nyord*, 'neologisms', but whose definition in this case differs slightly from many

other contexts (see further next section). The work related to new entries is a separate track in relation to the other two revision approaches, in that it does not involve the existing entries to any significant extent. The task is more straightforward in the way that the old, retagged structure does not have to be taken into account (except in the case of cross-references). Instead, the content of the entries can be written directly into the purpose-designed software tool.³

Overall, the addition of new entries is the one of the three tasks in which the most time will be invested. In the current project plan, it is estimated that about 50% of the resources (it should be said that the editorial staff currently consists of 17 people, with different employment rate) will be spent on this work, and the remaining 50% on the other two tasks described above (including IT developments).

The assignment of finding, collecting and defining these neologisms are described in more detail in the next section.

4. New Entries – and the Search for “Old Neologisms”

As has already been stated, new entries are needed in SAOB, and this need is not evenly distributed throughout the dictionary. The lack is even more obvious in the first volumes, or rather in the first half of the dictionary, partly because the excerption of source material to a large extent was done in advance, but also because several entries were published at the end of the 19th and at the beginning of the 20th century. Therefore, also in this part of the work, the focus is on words with initial letters from A–R. Based on these SAOB’s needs, a neologism in this context is not to be understood as a lexical unit perceived as new or recently established or not yet (fully) lexicalized (see definitions discussed by Klosa & Lungen, 2018, pp. 560–661; cf. also Klosa-Kückelhaus & Kernerman, 2023). The perspective here is slightly different since the concept of *new* is relative, and in the context of SAOB’s long time span, it aims at something else than, for instance, in the latest edition of a modern contemporary dictionary. To address this relativity, neologisms, or the new features, are here simply related to the lemma list of SAOB. It refers to a headword that the dictionary will be supplemented with, because it is considered a deficiency (due to SAOB’s aim to reflect the language in the most representative possible way) that it is missing in the first edition. A definition thus more in line with the one mentioned by Afentoulidou & Christofidou (2021, p. 599): “any lexical unit being classified as a member of the active vocabulary of the speech community [...], based on certain criteria, which would allow its inclusion in general-purpose dictionaries.” (More specific criteria are commented below.)

A relevant question to ask before the addition of new entries in SAOB’s revision project is how many headwords should be added. How much can be done, and what is a sufficient amount to make the dictionary better and to overcome the biggest flaws? There is, of course, no clear answer to these questions, but the editors have come to

³ The development and design of the application of the software will not be discussed in more detail in this paper. But it must be emphasized that the adaptation to the structure of SAOB is a huge effort that has been going on for several years.

the conclusion that 10,000 lemmas, corresponding to about one printed volume, is a reasonable number, and that is what is aimed for. The estimate is based both on how much of the total amount the supplement makes up, but of course also on how much time of work will be spent.

Although the first edition of SAOB contains a respectable number of entries, more than any other Swedish dictionary, it is not a difficult task to find relevant words to add to the dictionary. On the contrary, the challenge is to restrict the selection and to compile narrow criteria for inclusion.

4.1 Sources for Neologism Candidates and Specified Selection Criteria

First of all, it must be decided what sources to use to find neologism candidates. In the first edition of SAOB the framework was fairly fixed, and, more or less, the later editors came to a set table; an excerpted, physical source material⁴ was available and the words contained therein ended up in the dictionary mainly if they met the following criteria: they were not proper names, they were not exclusive technical terms or dialect words, and they were not purely occasional formations. The vast majority of simplex words in the material were thus included, in addition to a rich selection of compounds, derivatives and particle verbs. Now the situation is different, and input can be obtained from almost anywhere. The aim will still be to describe the Swedish language in as comprehensive a way as possible, and the criteria described above still apply. Five different channels for accessing neologism candidates have been narrowed and will be used, of which the work so far is based on one of them.

The first source of candidates is found in-house. Since the excerpting of the printed sources used for the first edition was executed continuously from the preface (mentioned above) to the 1990s on words in all the alphabet, though some of the letters were already treated in the dictionary, a certain part of the material labeled *late incoming* is available in the archive of the dictionary, which has not been utilized so far. This partial material consists of approx. 2–300,000 excerpts. For the sake of candidate selection, the lemmas in this material are being manually excerpted and listed. The compiled lemma list from this source material is later to be matched and aligned with the lemma list of SAOB. Samples show that relevant candidates can be found this way.

As mentioned in Section 2.2, SAOB is digitally available together with the two other dictionaries of the Swedish Academy, SO and SAOL, on the site <https://svenska.se/>. Also their lemma lists are matched with SAOB's. The result shows that about 40,000 lemmas found in SO and SAOL are lacking in SAOB within the relevant alphabetic span. The SAOB editors have gone through this list in pairs and have selected just over 3,000 relevant words to include in the revised version of the dictionary. This selection forms the basis of the work on new entries that has begun, and which at

⁴ SAOB's source material consists of approx. 8,2 million excerpts collected from about 20,000 sources of different genres. See e.g., Svensson (1992) for a discussion on SAOB's material with emphasis on the representativeness of the sources; cf. Malmgren (2011) for a shorter description in English.

the time of writing has arrived at the last words beginning with the letter B in the alphabetical order. (How the selection was carried out is described in more detail below.)

In addition, a targeted corpus has been compiled for the purpose of the dictionary. This corpus, called SAOB1950, is freely available on Språkbanken text's search tool *Korp* (for more information about this tool see e.g., Borin et al., 2012). The corpus contains material from fiction and non-fiction books collected from the university library in Lund with the intention of balancing the newspaper text bias that is present in other available (Swedish) corpora. The books are published, as the name of the corpus suggests, from 1950 onwards and they are selected to cover several different genres, both by direct selection (certain genres and subgenres) and by random selection (every *n*th book on every *n*th shelf). The corpus is relatively small, about 50,000,000 tokens, but of high quality for the dictionary purposes. The total amount of tokens in the corpus will be matched with SAOB's lemma list. Samples show that this source will also be useful. It contains a slightly different kind of words than the aforementioned sources.

Except for these sources, candidates are collected manually by listing lemmas that are relevant to add in connection with other neologisms being added. For example, the definition of the neologism compound *afterski band*: "that typically plays covers", resulted in the word *cover* ending up on the list of neologism candidates. In addition, candidates can be collected through the log files of so-called *not-found*s, i.e., frequently recurring non-successful searches of words that are not included in the dictionary (cf. Bäckrud et al., 2021). The point of having several different channels to collect candidates is, of course, that a more balanced influx and representation from different areas can be accomplished. However, in this case, more emphasis is placed on quality than quantity. What is decisive is the composition and different nature of the sources.

The following criteria (in decreasing order of priority) form the basis for the assessment of candidates.⁵ Words that are given priority:

- belong to (or in an earlier language stage have belonged to) almost all language users' (at least passive) vocabulary; for example (here and henceforth, if no translation is given: the Swedish equivalents of): *abstinence*, *acetone*, *adhd*, *airbag*, *activate*, *allergy*, *avocado*;
- can be considered somewhat specialized, often belonging to a more or less technical language, but which quite a few language users may need to acquire through practical needs or leisure interests or community involvement or encountering illness and the like; examples: e.g.: *agility*, *antonym*, *arbitrage*, *assist* ('scoring pass'), *audiologist*;
- may be considered "difficult" or "learned" words that should be explained in a dictionary (even if they have low frequency), e.g., *annihilate*, *adiaphora*, *arborist*;

⁵ The candidate collection based on SO's and SAOL's lemma lists was performed in pairs letter by letter. The words that both editors in the pair had selected ended up on the final list of neologism candidates.

- occur in the meta-text of SAOB (or are found in the list of lemmas as the second part of a compound); e.g.: *anion* (in the entry CATION), *anträffbar* ('accessible, available', in the entries O-ANTRÄFFBAR, SVÅR-ANTRÄFFBAR).

Simplex words are prioritized, not least since it is unproblematic for the editors during the actual editing process to supplement the simplex entry with one or more nest-lemmas if needed. But also (frequent) compounds and derivatives are collected if they are:

- lexicalized in an opaque special meaning (at least in one of its meanings), because the relationship between the parts is completely unique, or one or both parts as a simplex word exhibits such extensive polysemy that several different "meaning sums" are plausible for the compound; for example: *aha-experience*, *daghem* ('day care');
- more or less semantically transparent, but have nevertheless established themselves as "just what it is called" despite fully possible alternative expressions; that is, they are unpredictable from a production point of view; for example: *allergic shock*, *alliance free*, *atomic bomb*.

Other words are to be left unmarked as instances of a low-priority lack in the list of lemmas. However, even with the specified criteria, there is an imminent risk of the selection becoming too abundant. Therefore, some priority guidelines have been formulated. Firstly, names of varieties and sorts and name-like designations (which are not highly frequent or culturally central) are de-prioritized. This means for example that names of food dishes or all dog and cat breeds are not prioritized. Neither are all designations for inhabitants and languages collected. In fact, the aim is not to prioritize complete descriptions of different systems just for the sake of systematicity. A more pragmatic approach must be applied. Instead, emphasis is placed on frequency, relevance and cultural centrality. Secondly, the fact that SAOB has a historical claim is highlighted. In other words, slightly older words that are at risk of being phased out from the purely contemporary dictionaries are given high priority in SAOB. This means that words such as *bakelite*, denoting a material that is hardly used nowadays, *zairisk* ('Zairian'), despite being a designation for nationality, but in the capacity of a no longer officially valid term, the above-mentioned *daghem*, officially replaced by *förskola* ('preschool'), and the 1970s-scented *barkök* ('kitchen bar'), will end up high on the list. Thus, the task to large extent is to search for "old neologisms".

4.2 Editing and Reviewing

The work on adding new entries has at this point made some progress. So far, words like *acne*, *app* and *baseball* and such previously mentioned words as *airbag*, *activate*, *allergy*, *bakelite*, and *barkök* etc. have been added. (However, these entries have not yet been published online – a first round is expected to be published in 2025.) But it is not only the headwords themselves that are new, also the principles for describing the words and reviewing the written descriptions have several new elements. In this section, this news will be briefly discussed.

As mentioned in the previous section, there are a number of channels to identify neologism candidates. In the analysis and description of the individual word, however, the editors use more resources than these. As highlighted in Section 2.1, the first quote or sample of a form or meaning is important in SAOB. To access these, the editors use several different corpora plus the physical material available at the editorial office. Of these corpora, special mention can be made of the newspaper database of the National Library of Sweden, *Svenska Tidningar* (available at: <https://tidningar.kb.se/>), where a first quote is often found. In this respect, of course, SAOB's own corpus SAOB1950 is also important. In connection with the previously mentioned revision tour, many comments were received about the complexity of SAOB's definition text. As mentioned in Section 3.2, systematic work is being done to make the hierarchy in the definitions clearer. In addition, the neologism editors now have the opportunity to express themselves in a more accessible way in the definitions, which, for example, means that they can now avoid long (preceding) noun phrases (in favor of relative clauses) and complicated structural solutions, for instance, requiring parentheses within parentheses. In addition, typicality of uses and attributes, etc. can be made explicit with expressions such as: *typically of X or Y*, or *for instance in terms of Z* etc.

The editors are currently working with a span of 25 neologisms at a time. Information about all sources and citations used in each span of entries are saved in separate documents. When the first draft is completed, a colleague reviews the entries for all content (including cited sources). After further processing, the entries are reviewed by two other colleagues, before it is proofread and ready for publication. At the time of writing, the work has not yet reached this point, but with the preparatory work done, the route is thus marked out.

5. Concluding Remarks

In sum, SAOB is facing both exciting and challenging times. Everything can be questioned and much can be improved, and the possibilities are in many respects unlimited, for better or worse. It is therefore important to take a pragmatic approach, and the editors must in some respects apply a trial-and-error method in the initial revision work. After fairly thorough preliminary efforts, however, there is a stable basis to start from, and the fact that the dictionary will henceforth be continuously updated implies that mistakes, errors and inadequacies can be corrected step by step as they are discovered.

The various parts of the revision are made clear and manageable to an increased extent, not least when it comes to work with neologisms. Obvious gaps and deficiencies in the lemma list are identified, and the intention is not to aim at so-called comet words or the very latest novelties in the vocabulary, there are better dictionaries for that purpose. It thus seems appropriate for a diachronic-contemporary dictionary like SAOB to set the sights on old neologisms.

References

- Afentoulidou, V., & Christofidou, A. (2021). It's a long way to a dictionary: Towards a corpus-based dictionary of neologisms. In Z. Gavrilidou, L. Mitits, & S. Kiosses (Eds.), *Lexicography for inclusion: Proceedings of the XIX Congress of the European Association for Lexicography* (pp. 597–606). Democritus University of Thrace.
- Arts, T. (2014). The Making of a Large English-Arabic/Arabic-English Dictionary: the Oxford Arabic Dictionary. In A. Abel, C. Vettori, & N. Ralli (Eds.), *Proceedings of the XVI EURALEX International Congress: The user in focus* (pp. 109–123). Eurac.
- Borin, L., Forsberg, M., & Roxendal, J. (2012). Korp: The corpus infrastructure of Språkbanken. In N. Calzolari, K. Choukri, T. Declerck, M. Uğur Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of LREC 2012* (pp. 474–478). ELRA.
- Bäckerud, E., Nilsson, P., & Sköldberg, S. (2021). Så används Svenska Akademiens ordböcker på nätet. Implicit och explicit feedback från användarna. In C. Sandström, U.-M. Forsberg, C. af Hällström-Reijonen, M. Lehtonen, & K. Ruppel (Eds.), *Nordiska studier i lexikografi 15* (pp. 91–101). Nordiska föreningen för lexikografi.
- DWB: *Deutsches Wörterbuch von Jacob Grimm und Wilhelm Grimm, digitalisierte Fassung im Wörterbuchnetz des Trier Center for Digital Humanities, Version 01/23*. Retrieved July 27, 2024, from <https://woerterbuchnetz.de/?sigle=DWB#0>
- Eaker, B. (2006). Historisk lexikografi i praktiken. Om momentordningen i SAOB. *LexicoNordica*, 13, 41–54.
- Elliott, L., & Williams, S. (2006). Pasadena: a new editing system for the Oxford English Dictionary. In E. Corino, C. Marelllo, & C. Onesti (Eds.), *Atti del XII Congresso internazionale di lessicografia/ Proceedings of the XIIth Euralex International Congress* (pp. 257–264). Edizioni dell'Orso.
- Hast, S. (1983). Om SAOB:s material. *Nysvenska studier*, 6, 159–192.
- Kinable, D. (2006). Computerized restoration of historical dictionaries: uniformization and date-assigning in dictionary quotations of the Woordenboek der Nederlandsche Taal. *Journal of literary and linguistic computing*, 21(3), 295–310.
- Klosa-Kückelhaus, A., & Kernerman, I. (2023). Unregistered words, neologisms, and dictionaries: An introduction. *Lexicography: Journal of Asialex*, 10(2), 87–93.
- Klosa, A., & Lungen, H. (2018). New German Words: Detection and Description. In J. Čibej, V. Gorjanc, I. Kosem, & S. Krek (Eds.), *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts* (pp. 559–569). Ljubljana University Press.
- Kocjančič, P., Krek, S., & Climent, P. (2006). Bilingual Dictionaries and IDM DPS: The Development of a Corpus-driven Slovenian-English Pocket Dictionary and its Implementation in the IDM Dictionary Publishing System. In E. Corino, C. Marelllo, & C. Onesti (Eds.), *Proceedings of the XII EURALEX International Congress* (pp. 431–438). Università di Torino.
- Larsson, L. (2014). En “mer l. mindre stor” stor ordbok – om variationerna i SAOB:s omfång och ambitionsnivå. *LexicoNordica*, 21, 61–80.

Lundbladh, C.-E. (2003). Kritiken av SAOB. *LexicoNordica*, 13, 273–284.

Malmgren, S.-G. (2011). Das schwedische Akademiewörterbuch und das Wörterbuch der dänischen Sprache). In U. Haß (Ed.), *Grosse Lexika und Wörterbücher Europas. Europäische Enzyklopädien und Wörterbücher in historischen Porträts* (pp. 301–315). de Gruyter.

Nilsson, P. (2023). *Lexikal betydelseutveckling i teori och praktik: en analys av fem definitionsformler i Svenska Akademiens ordbok och de semantiska förändringsmekanismer som de beskriver*. [Doctoral dissertation, Lund University].

Nilsson, P. (2024a). Figurative, transferred or extended use? The use of semantic labels in the first edition and the revised version of the Swedish Academy Dictionary. In K. Š. Despot, A. Ostroški Anić, & I. Brač (Eds.), *Lexicography and Semantics. Proceedings of the 21st EURALEX International Congress* (pp. 123–131). Institute for the Croatian Language.

Nilsson, P. (2024b). Glancing back, looking forward: Swedish Academy Dictionary completed after 130 years – now time for revision. [Manuscript in preparation]

ODS: *Ordbog over det danske Sprog*, grundlagt af Verner Dahlerup, udg. af Det Danske Sprog- og Litteraturselskab, bd 1–28, København 1918–1956. Retrieved July 29, 2024, from <https://ordnet.dk/ods>

OED: *The Oxford English Dictionary*, second edition, vol. I–XX (of the corrected reissue (1933) with an Introduction, Supplement, and Bibliography of *A New English Dictionary on Historical Principles*, vol. I–X, ed. by James A. H. Murray et al. (1884–1928) and Supplement, vol. I–IV, ed. by R. W. Burchfield (1972–1986)), Oxford 1989. Retrieved July 29, 2024 from <https://www.oed.com>

OSM: Söderwall, K.-F. (Ed.) (1884–1918). *Ordbok öfver det svenska medeltids-språket*. Retrieved July 29, 2024, from <https://spraakbanken.gu.se/resurser/soederwall>

SAOB: *Svenska Akademiens ordbok* (1898–2023). Glerups. Retrieved July 29, 2024 from <https://www.saob.se/> and <https://svenska.se/>

SAOL: *Svenska Akademiens ordlista* (14th edition). (2015). Retrieved July 29, 2024 from <https://svenska.se/>

Scheider, M. (2001). Die Neubearbeitung des ‚Deutschen Wörterbuchs‘ von Jacob Grimm und Wilhelm Grimm. *Zeitschrift für Germanistik*, 11(3), 600–607.

Simpson, J. (2020). *Preface to the Third Edition of the OED*. Retrieved July 29, 2024 from <https://www.oed.com/information/about-the-oed/history-of-the-oed/preface-to-the-third-edition-of-the-oed/?tl=true>

SO: *Svensk ordbok utgiven av Svenska Akademien* (2nd edition). (2021). Retrieved July 29, 2024, from <https://svenska.se/>

Svensén, B. (2009). *A handbook of lexicography: the theory and practice of dictionarymaking*. Cambridge University Press.

Svensén, B. (2015). *Snille och smak. Svenska Akademien förr och nu*. Norstedts.

Svensson, L. (1992). Om SAOB:s material och en bibliografisk databas. In R. Vatveldt Fjeld (Red.), *Nordiske studier i leksikografi 1. Rapport fra konferense om leksikografi i Norden* (pp. 369–379). Nordiska föreningen för lexikografi.

Statutes of the Swedish Academy (1786). Retrieved July, 29, 2024, from <https://sa-admin.lb.se/assets/3c46295e-685c-4bf1-a39d-61f1d8cd6f85.pdf>

Teleman, U. (2015). *Nordister och nordistik i Lund – från 1860- till 1930-tal*. Lundastudier i nordisk språkvetenskap A 73. Lund.

Trap-Jensen, L. (2019). State-of-the-art on monolingual lexicography for Denmark (Danish). *Slovenščina* 2.0, 7(1), 1–12. <http://dx.doi.org/10.4312/slo2.0.2019.2.1-16>

WNT: *Woordenboek der Nederlandsche Taal*, deel I–XXIX + Bronnenlijst (1943, 1966) + Supplement I A–Azuursteen (1956), 's-Gravenhage & Leiden 1882–1998. Retrieved July 29, 2024, from <https://gtb.ivdnt.org/search/?owner=wnt>

Contact information

Pär Nilsson

Svenska Akademiens ordboksredaktion
par.nilsson@svenskaakademien.se

Sanni Nimb, Nathalie C. H. Sørensen, and Jonas Jensen

MAKING DANISH THESAURUS DATA AVAILABLE TO RESEARCHERS

The WebDDB project

Abstract This study presents a project aiming to make thesaurus data available under an academic licence. The project is based on the printed thesaurus *Den Danske Begrebsordbog* (DDB) which covers approx. 80% of the Danish dictionary DDO (ordnet.dk/ddo). It presents more than 100,000 different words and expressions categorised and ordered semantically in 22 thematic chapters, and 888 named sections. The data is now downloadable at a webpage where it can be supplemented with different types of lexical information from other resources of choice, e.g., information on valency, etymology, or ontological type. The supplementation is possible due to shared sense id-numbers between the lemmas in the digital thesaurus manuscript, the Danish online dictionary DDO, the semantic lexicon COR.SEM, and a WordNet (DanNet). The webpage allows for new types of studies of the Danish vocabulary with semantic similarity as the starting point. As part of the project, more lemmas from the DDO were added to the digital manuscript which today covers 95% of the dictionary. The vocabulary as well as certain sections and lemmas denoting nationality, sexual orientation, gender identity etc. are thoroughly revised due to the change of attitudes towards this vocabulary in the last decade.

Keywords thesaurus; linked data; semantics; webpage

1. Introduction

The print version of the Danish thesaurus *Den Danske Begrebsordbog* (Nimb et al., 2014, Nimb et al., 2014/2015) covers of approx. 80% of the Danish dictionary DDO (ordnet.dk/ddo). It contains more than 100,000 different words and expressions categorised in 22 thematic chapters (e.g., ‘Arts and Culture’ and ‘Sports and Leisure’), and 888 named sections (e.g., ‘Films and Cinema’ and ‘Football’). In Figure 1., a selection of nouns from ‘Films and Cinema’ is shown.



Fig. 1: Nouns from the section Films and cinema: “**entertainment movie**, drama • **comedy**, farce, folk comedy, sitcom • **romance**, melodrama, tearjerker • **action movie**, high-octane movie, violent film, karate film, disaster film, war film • road movie • **agent thriller**, spy film • **crime film**, detective film • **thriller**, suspense drama, espionage thriller • **gangster film**, mafia film • **horror**, horror film, slasher, slasher film, gore, vampire film • **science fiction film**, sci-fi film, fantasy, adventure film • **western**, western movie, western film, oat opera, spaghetti western • **family film**, children’s film, teen movie, adult film • **animated film**, animation, computer animation, puppet film, cartoon, cartoon series

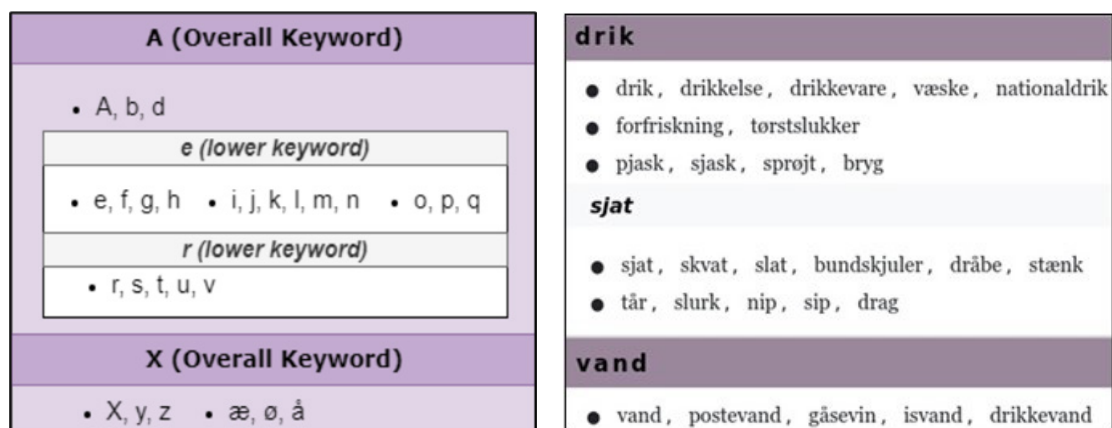


Fig. 2: The structure of the DDB thesaurus is shown to the left. Each letter represents a word. To the right, the nouns *drik* ('beverage', overall keyword), *sjat* ('small amount of a beverage', lower keyword), and *vand* ('water', overall keyword), all three followed by groups of near-synonymous nouns, illustrate the structure

The words are listed in semantic order in groups of persons, artefacts, acts, properties etc., some of which are initiated by a highlighted word which thereby functions as a keyword and headline to the following vocabulary, see Figure 2. The hierarchical structure provides valuable formalised knowledge of the varying degrees of semantic similarity between senses.

The structured and semantically ordered DDB data has proved useful in several internal projects due to links at sense level from the thesaurus to the DDO. In section 3 we describe the linked data and use of the thesaurus in different tasks and, in section 4, the new webpage. The next section focuses on the extension and review of the thesaurus manuscript before the online publication.

2. Review of the DDB Thesaurus 10 Years After: Additions and Outdated Section Vocabulary

Following the 2014/15 publication of the print version of the DDB, more than 20,000 words and expressions have been added to the digital manuscript. Within the same time span, language attitudes have changed considerably due to an increased focus on equality in society, having the consequence that in some cases the vocabulary, usage labels, keywords and section names have become outdated. To secure the quality and integrity of the data on the webpage we therefore took the opportunity to review the manuscript. This review brought attention to a number of larger issues, e.g., the discussion of whether to include obsolete words such as *dragesæd* (a rare term of ancient Greek origin meaning 'hostility' or 'conflict') and whether *plantemælk* (~plant milk) would be an appropriate keyword following a 2017 ruling by the Court of Justice of the EU which prohibits the use of the word 'milk' in the advertising and marketing of purely plant-based products. In the first case, *dragesæd* has been removed. We aim at covering only modern Danish corresponding to the DDO vocabulary (1950 until today). In the latter case, the synonym *plantedrik* ('plant-based drink') has been upgraded to keyword instead. In the thesaurus, we prefer neutral keywords, i.e., politically correct and non-derogatory keywords, whenever possible. The reviewers

were instructed to scour existing sections for potential incongruities, e.g., instances in which derogatory words or racial slurs featured next to neutral or unmarked terms denoting nationality, sexual orientation, gender identity, etc. In one such case, *trans*, a derogatory word denoting ‘crossdresser’, has been removed from the section ‘Dislike, antipathy’) and is now only part of the section ‘Clothes’. Likewise, the word *mobbeoffer* (‘victim of bullying’) was removed from the section ‘Ridiculous’ where it was grouped with derogatory words such as *ffog* (‘fool’), *nar* (‘jester’), *jubelidiot* (‘jubilant idiot’), and *klovn* (‘clown’). On the webpage, it is only presented in the section ‘Sarcasm, mockery’. Furthermore, a few section titles have been changed in the new version of the thesaurus data, e.g., the section ‘Ligestilling’ (‘Gender Equality’) is renamed ‘Ligestilling, diskrimination’ (‘Gender Equality, Discrimination’).

3. Linked Data

All words in the DDB thesaurus are linked to the corresponding DDO sense. The linked thesaurus and dictionary data have proven useful in several respects.

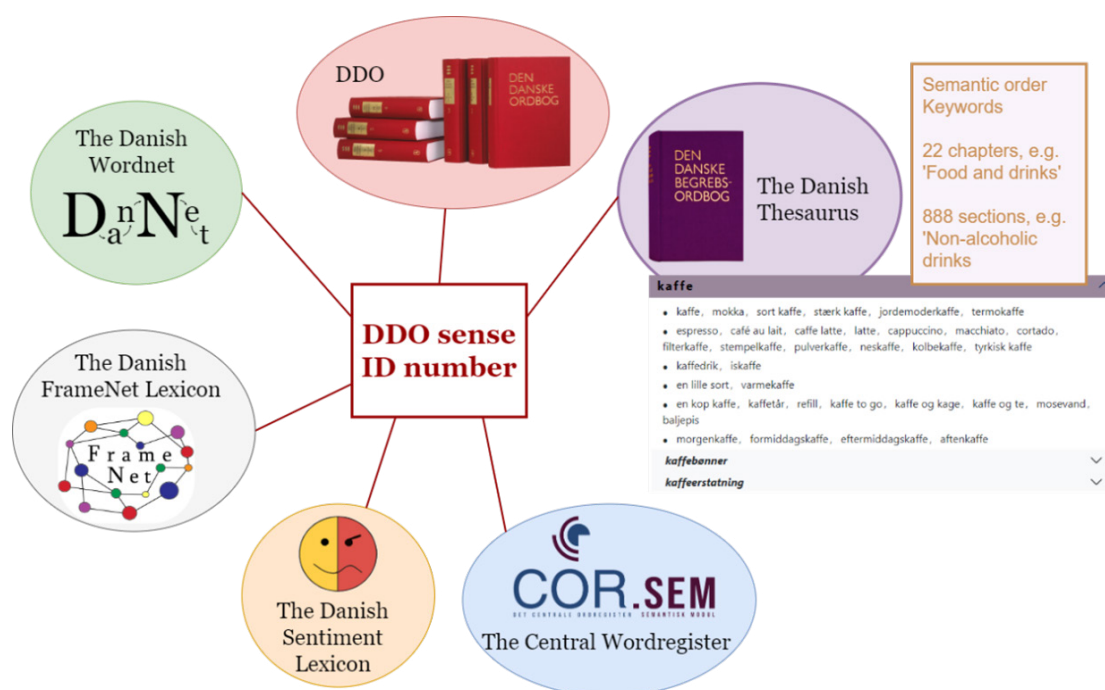


Fig. 3: The linked dictionaries and lexicons. The DDO dictionary and the DDB thesaurus are compiled by DSL and DanNet, the FrameNet lexicon, the sentiment lexicon and COR.SEM by DSL and CST, the University of Copenhagen

It allowed for the compilation of a FrameNet lexicon, based on thematic sections and semantic groups in the thesaurus combined with the valency patterns in the DDO which were assigned frames from Berkeley FrameNet (Nimb et al., 2017; Baker et al., 2003). A sentiment lexicon was developed from the vocabulary in negative and positive DDB sections combined with DDO information on style labels (Nimb et al., 2022). Recently the COR.SEM lexicon has been compiled based on the links (Pedersen et al., 2021; Nimb et al., 2024). Altogether, the resources constitute a lexical network at sense level which also includes the Danish WordNet DanNet (Pedersen

et al., 2009), see Figure 3. The links between the thesaurus and DDO are also used in the online DDO to present extracts of thesaurus section data based on the structure and keyword scope (Nimb et al., 2018). Recently, datasets based on the thesaurus have been used as LLM evaluation data as part of a Danish semantic reasoning benchmark (Pedersen et al., 2024). Based on the thesaurus where the closest word is the synonym (neighbour word), the next words are from the same section or chapter, and the word furthest away is from a different chapter, the task is to select the correct synonym or to identify the outlier out of 4 words (see also Nielsen & Hansen, 2017, and Camacho-Collados & Navigli, 2016).

4. The Webpage

The thesaurus data is available on a purpose-built webpage webddb.dsl.dk to be published in December 2024. We have used the semantic links between the DDB manuscript and the lexical resources to precompile a selection of datasets which can be combined with the selected vocabulary and downloaded under a research licence. Prior to this, a group of researchers specified the combined data which would be relevant in their respective research projects. For instance, linguists were interested in semantic groups combined with the DDO information on derivation patterns of the lemmas (verbs and the corresponding verbal nouns) to be able to study whether semantics influence patterns. Literary researchers showed interest in thesaurus section names for the purpose of automatic topic detection. In addition, formalised data (e.g., the thesaurus structure linked to additional formal information such as ontological type) was of interest to data scientists aiming to carry out automatic sense disambiguation. Format and other technical issues were also discussed.

On the webpage, researchers can access the data either by searching for words with a query string or by selecting a chapter and then a section to retrieve all words within that section. The data is shown as a table with the default columns: word form, chapter, section, overall keyword, lower keyword, and Part of Speech (see Figure 4).

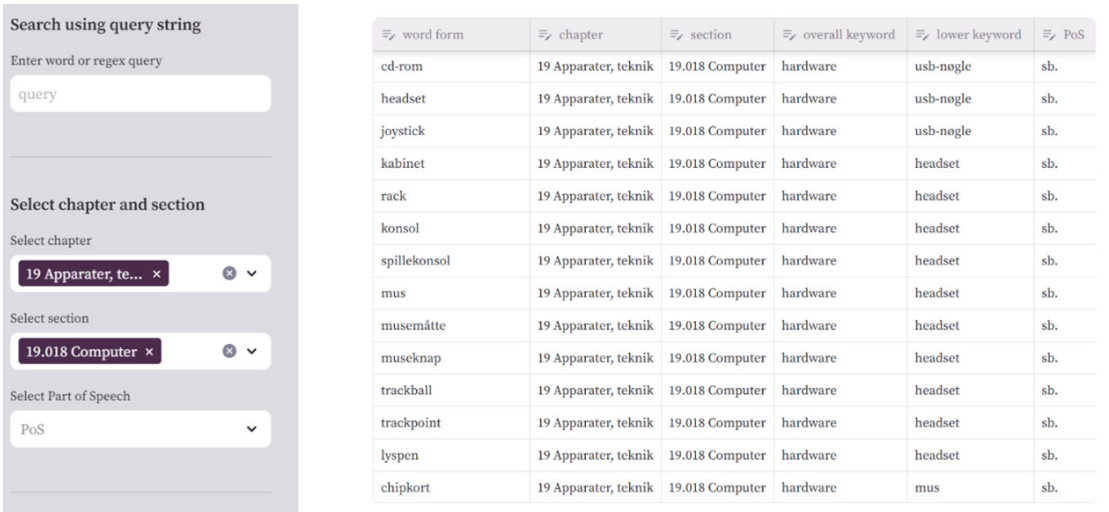


Fig. 4: The search interface, webddb.dsl.dk
Researchers can also select additional columns with information retrieved from either

the DDO or COR.SEM. Furthermore, each column can be filtered for specific values so that researchers can easily select a subset of the data - for instance by filtering for verbs in PoS and then a specific valency pattern. The available information is shown in Table 1.

Examples of filtered subsets of data are near synonym verbs in DDB with the sense ‘to think’, combined with information in DDO on which language they are borrowed from: *ræsonnere* (‘to reason’): French and Latin; *deducere* (‘to deduce’): Latin; *konkludere* (‘to conclude’): Latin; *tænke* (‘to think’): Middle Low German; *forestille sig* (‘to imagine’): Low German or German.

Table 1: Information types to be added to thesaurus groups of your own choice, webddb.dsl.dk

Information type	Source	Example
Grammar: valency/ construction pattern	DDO	verb <i>smage</i> (‘to taste’): NGN <i>smager</i> NGT (‘somebody tastes (something)’); NGN <i>smager på</i> NGT (‘somebody tastes Ø something’)
Etymology: borrowed from which language(s)?	DDO	verb <i>smage</i> (‘to taste’): nedertysk (‘low German’) <i>smaken</i>
Etymology: dating (year)	DDO	noun <i>computer</i> : 1957
Morphology: Derivation patterns	DDO	verb <i>smage</i> : noun <i>smager</i> ‘a taster’, noun <i>smagning</i> (‘tasting’)
Sentiment value	COR.SEM	verb <i>lugte</i> ‘to smell’: ‘-1’, negative polarity
Ontological type	COR.SEM	verb <i>smage</i> (‘to taste’): Experience+Physical
Semantic frame (values based on Berkeley FrameNet)	COR.SEM	verb <i>smage</i> (‘to taste’): frame=Tasting
Synset in the Danish WordNet	DanNet	https://wordnet.dk/dannet/data/synset-72897

Another example is near synonymous words for the noun *computer* (‘computer’) combined with information on when the words first appeared in Danish (provided that this information is available in DDO): *datamaskine*: 1953; *elektronhjerne* (lit. ‘electron brain’): 1954; *computer*: 1959; *datamat*: 1966; *terminal*: 1968; *lommeregner* (lit. ‘pocket calculator’): 1970; *pc*: 1984; *laptop*: 1990; *client/server*: 1991; *desktop*: 1993. Lastly, the table shown on the webpage can also be downloaded as a tsv-file to save specific searches for further studies and applications.

5. Conclusion

The thesaurus webpage will be accessible to researchers in December 2024. Access to the lexical data under an academic licence allows for a wide range of new applications in the research community, either based on standalone thesaurus data or combined with data from DDO and formal semantic lexicons. DSL is also going to publish a second version of the thesaurus in 2025/2026 based on the extended and revised manuscript. Unlike the print version from 2014, this will be published in the form of an open access online thesaurus linked to DDO and containing approx. 20% more lemmas and senses.

References

- Baker, C. F., Fillmore, C. J., & Cronin, B. (2003). The structure of the FrameNet database. *International Journal of Lexicography*, 16(3), 281–296.
- Berkeley FrameNet. Retrieved May 27, 2024 from <https://framenet.icsi.berkeley.edu/>
- Camacho-Collados, J., & Navigli, R. (2016). Find the word that does not belong: A Framework for an Intrinsic Evaluation of Word Vector Representations. In *Proceedings of the 1st Workshop on Evaluating Vector Space Representations for NLP* (pp. 43–50). Association for Computational Linguistics.
- DanNet. Retrieved May 27, 2024 from <https://wordnet.dk/DanNet>
- Den Danske Ordbog (DDO) (2009-). Online ordnet.dk/ddo. Det Danske Sprog- og Litteraturselskab
- Det Danske FrameNet-leksikon. Retrieved May 27, 2024 from <https://korpus.dsl.dk/resources/details/framenet.html>
- Det Danske Sentiment-leksikon. Retrieved May 27, 2024 from <https://github.com/dsldk/danish-sentiment-lexicon>
- Hjorth, E., & Kristensen, K. (Eds.). (2003–2005). *Den Danske Ordbog*, Society for Danish Language and Literature & Gyldendal.
- Nielsen, F. Å., & Hansen, L. K. (2017). Open semantic analysis: The case of word level semantics in Danish. In *Human Language Technologies as a Challenge for Computer Science and Linguistics* (pp. 415–419).
- Nimb, S., Braasch, A., Olsen, S., Pedersen, B. S., & Søgaard, A. (2017). From Thesaurus to FrameNet. In *Electronic Lexicography in the 21st century: Proceedings of eLex 2017 conference* (pp. 1–22).
- Nimb, S., Flörke, I., Olsen, S., Pedersen, B. S., & Sørensen, N. C. H. (2024). COR.SEM, a new formal semantic lexicon for Danish. In K. Š. Despot, A. Ostroški Anić, & I. Brač, *Proceedings of the 21st EURALEX International Congress, Lexicography and Semantics* (this issue).
- Nimb, S., Lorentzen, H., Troelsgård, T., & Theilgaard, L. (2015). *Den Danske Begrebsordbog*. Det Danske Sprog- og Litteraturselskab
- Nimb, S., Olsen, S., Pedersen, B. S., & Troelsgaard, T. (2022). A Thesaurus-based Sentiment Lexicon for Danish: The Danish Sentiment Lexicon. In *Proceedings of the Language Resources and Evaluation Conference: LREC2022* (pp. 2826–2832). Marseille: European Language Resources Association.
- Nimb, S., Sørensen, N. H., & Troelsgård, T. (2018). From standalone thesaurus to integrated related words in the Danish Dictionary. In J. Čibej, V. Gorjanc, I. Kosem, & S. Krek (Eds.), *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts* (pp. 915–923). Ljubljana: Ljubljana University Press, Faculty of Arts.

Nimb, S., Trap-Jensen, L., & Lorentzen, H. (2014). The Danish Thesaurus: Problems and Perspectives. In A. Abel, C. Vettori, & N. Ralli (Eds.) *Proceedings of the XVI EURALEX International Congress: The User in Focus* (pp. 191–199).

Pedersen, B. S., Nimb, S., Asmussen, J., Sørensen, N., Trap-Jensen, L., & Lorentzen, H. (2009). DanNet – the challenge of compiling a WordNet for Danish by reusing a monolingual dictionary. In *Language Resources and Evaluation, Computational Linguistics Series*. 10.1007/s10579-009-9092-1. 31 s.

Pedersen, B. S., Sørensen, N. C. H., Nimb, S., Flörke, I., Olsen, S., & Troelsgård, T. (2022). Compiling a Suitable Level of Sense Granularity in a Lexicon for AI Purposes: The Open Source COR-Lexicon. In *Proceedings of the Language Resources and Evaluation Conference: LREC2022* (pp. 51–60). Marseille: European Language Resources Association.

Pedersen, B. S., Sørensen, N. C. S., Olsen, S., Nimb, S., & Gray, S. (2024). Towards a Danish Semantic Reasoning Benchmark – Compiled from Lexical-Semantic Resources for Assessing Selected Language Understanding Capabilities of Large Language Models. In *Proceedings of LREC-COLING 2024*, Torino, Italy.

Contact information

Sanni Nimb

Society for Danish Language and Literature (DSL)
sn@dsl.dk

Nathalie C. H. Sørensen

Society for Danish Language and Literature (DSL)
nats@dsl.dk

Jonas Jensen

Society for Danish Language and Literature (DSL)
jj@dsl.dk

Christian-Emil Smith Ore and Oddrun Grønvik

THE SPOKEN WORD AS REPRESENTED IN NORSK ORDBOK

Abstract Spoken language is the prerequisite of written standard languages for living language communities. Yet written sources dominate lexicographic description of standard languages, and awareness of dictionaries that specifically source speech seems limited. In *Norsk Ordbok* (*The Norwegian Dictionary*), and in the Language Collections on which the dictionary is based, oral materials are perceived as input to the national language Nynorsk, written or spoken. The purpose is integration into one whole, not a series of parallel lexical registers. Legitimacy is aimed at by explicit sourcing of linguistic information, whether from speech or literature. This paper looks at how speech is sourced within the entries of the dictionary *Norsk Ordbok* (NO), particularly at the sourcing of definitions. Explicit sourcing of speech in connection with definitions facilitates investigating the contribution of speech materials to *Norsk Ordbok* as a whole, and if and how the differences between speech and written text is reflected in *Norsk Ordbok*.

Keywords speech; dialect; vernacular; standard language; location; location source

1. Introduction

Speech-based dictionaries range from major multi-volume scholarly dictionaries (cf. *Dictionary of Finnish Dialects*, 2024, the *English Dialect Dictionary*, Wright, 1898–1905), to spontaneous web creations. A web search for “dialect lexicography” gives ca. 4 000 matches. Some ambitious modern digital dialect dictionaries are discussed in Euralex contributions (cf. Tier & Keumeulen, 2010). Yet speech sourced lexicography is generally seen as something quite different from mother tongue lexicography concerning a standard language.

Norsk Ordbok (NO) is an exception. The full title of NO is *Norsk Ordbok. Ordbok over det norske folkemålet og det nynorske skriftmålet* ‘The Norwegian Dictionary’. Dictionary of the Norwegian vernacular and the Nynorsk written standard’. The aim of integrating the documentation of speech and writing motivates the editorial system of NO, from lemma selection to the details of source sorting and listing (Skard, 1932; Hellevik, 1956), but until recently, it has been impossible to examine the whole in order to see how well the result meets the original requirements.

This has to do with a change in production method. The first (paper) edition of NO was produced in two stages. The alphabetical section *a-h* was edited on paper in the period 1946–2002. The alphabetical section *i-å* was edited into a relational database under a consistent set of editorial rules in the project NO2014 (2002–2015). The last eight volumes were produced from this editorial system.

In 2016, NO and the Norwegian Language Collections (Grønvik, 2015, p. 32) were moved from the University of Oslo to the University of Bergen. In 2019, the planned revision of the alphabetical section *a-h* started. The editorial system used for the alphabetical section *i-å* is still in use (though undergoing rewriting). The alphabetical section *a-h* of the printed dictionary is not complete in the dictionary database. Therefore, the editorial database is not well suited for analysing NO as a whole. Fortunately, the text of first two volumes were retro-digitized as formatted text files in 2004, and the editors' files for the rest of the alphabet section *a-h* are available as text files with a simple field mark-up. All the files have been thoroughly analysed and a given a semi-automatic xml mark-up at the level of detail found in the editorial system. This was first done in 2004 and then revised in 2023. As the alphabetical section *a-h* is under revision, the first edition has recently been published separately on the web as *Norsk Ordbok AH 2005* and *Norsk Ordbok IÅ 2016*. The latter is based on an xml-export from the editorial system of the alphabetical section *i-å* as it was in 2016. Both have been included in the extensive collection of background material underlying the revision of *a-h*.

NO can be accessed in several ways. It exists as a printed dictionary in 12 volumes. The editorial system is accessible for editors and others with special interests in the material. NO also exists as a web version that takes data directly from the editorial system, giving the current state of the contents. In addition, there are the fixed version *Norsk Ordbok AH 2005* and *Norsk Ordbok IÅ 2016*.

The whole of NO is now available for consistent analysis, and it is time to look at how NO fulfils its original, ambitious aim. The analyses referred to in this paper are based on the xml-texts for *a-h* and the editorial system (a relational database) for *i-å*.

2. The Terms 'Dialect' and 'Location'

The definition of the term 'dialect' may vary from one language community to another. In the study of the Norwegian spoken language, the following definitions are used:

Dialect = a geographically delimited language system (with a specific and defined phonology and morphology, to a lesser degree syntax and lexicon) (Sandøy, 1985, p. 68)

Norway lacks a spoken standard language ("received pronunciation") with the accompanying status and rules for usage that many other language communities have. In contrast, the space for and social acceptance of dialects is considerable (Netland & Opsahl, 2024). The closest one gets to a standard for speech is spoken versions of the two written standards Bokmål or Nynorsk, as used by hosts and announcers in news programmes and programme information in the national broadcasting corporation NRK. The varieties "spoken Bokmål or Nynorsk" covers word forms only, not intonation or accent. "In other programmes and in news items (reportage, commentaries, reports, interviews etc.) dialect may be used, including in national broadcasting" (NRK språket, 2007)

It follows that NO attempts to describe the Norwegian vernacular and the Nynorsk standard language in speech and writing, without assigning a limited register of genres or styles to the one or the other. Sections 6–8 deals with this issue in more detail. In this paper, “speech” is used in parallel with “dialect” in referring to dictionary entries built on oral materials.

Location = named place or area (where a dialect is used). When locations are explicitly stated in an NO definition, it can be assumed that the Language Collections at the University of Bergen hold documentation of a spoken usage of the word defined.

There are two referencing systems in use for referring to a place with regard to dialects, landscapes and (the area of) local municipalities. Landscape location is the older system, based on former administrative districts. Municipality location, referring to the lowest level in the administrative system, has been the preferred reference system for the study of the Norwegian vernacular since before World War II. The municipalities were based on the parishes as they were in 1837. This administrative partitioning of the country is consistent with the dialect borders. In the Language Collections at the University of Bergen, both location systems are used, but municipalities dominate in the more recent collections.

The use and acceptance of dialects in formal contexts in Norwegian society is accompanied by a strong interest in standard language and dialects among language users. It is important for NO to locate speech information through the entries, as users look for this information.

The location hierarchy of NO is a synthesis of the locations used by Norwegian dialectologists from 1870 and onwards. The oldest term, “landscape”, is in the middle (Table 1, level 4).

NO differs from previous Norwegian dictionaries in giving very detailed information about the location of usage. For example, the linguist Ivar Aasen in his dictionary (Aasen, 1873) did not go below “landscape”, which gave him a list of less than 200 possible locations, while NO uses about 900 locations, mostly the 750 municipalities (*kommune*) as they were in the 1940s, the most fine-grained administrative organisation of Norway at any time.

Table 1: The *Norsk Ordbok* location hierarchy

Level	Norwegian term	Number	English	Example
1	Kommune	750	Municipality	Nordre Land
2	Tvillingkommune	16	Adjoining municipalities with a joint name	Land
3	Del av landskap	6	Part of landscape	Nedre Romerike
4	Landskap	58	Landscape	Romerike
5	Del av fylke	5	Part of county	Vest-Oppland
6	Fylke	19	County	Oppland
7	Del av landsdel	6	Part of province	Flatbygdene på Austlandet
8	Landsdel	5	Province	Austlandet
9	Region	2	Region	Nordafjells
10	Land	1	Country	Noreg

In the electronic editorial system of NO, the location system is organized as a hierarchy of ten levels, starting from below with “kommune” (‘municipality’) and ending in the country as a whole “Noreg” (‘Norway’), as shown in Table 1. When half or more units at one level are registered, editors can generate the higher level automatically. The sequence of locations is fixed, starting in the Southeast and ending in the North next to the border shared with Russia.

The location hierarchy is linked to a digital map of Norway with the municipal borders valid in the 1940s, and the locations of word forms, senses and usage examples can be shown in maps as in Figure 1.

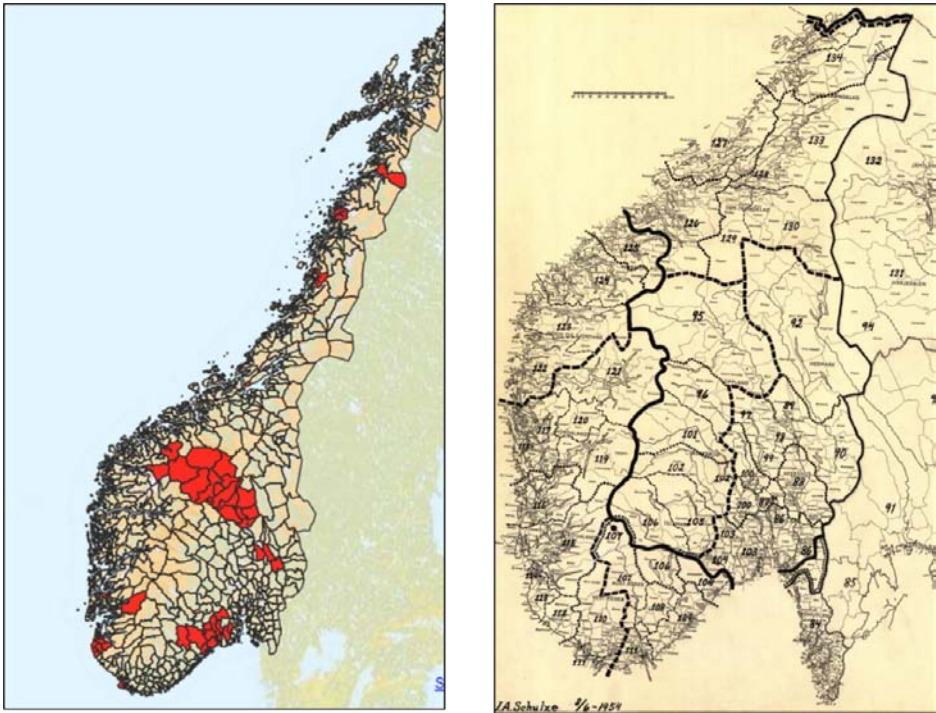


Fig. 1: Location sources (based on municipalities in 1947) for the NO entry IV *skrella* v. sense 3 ‘remove peel (from something)’

Fig. 2: Map ND001 from the Norwegian Dialect Atlas, showing the chief isoglosses for the dialect landscapes of Southern Norway, the East-West divide (thick black line) being the most important one

3. Editorial Rules in Norsk Ordbok for the Documentation of Speech

3.1 Origins

The written standard Nynorsk is based on the ground-breaking work by the Norwegian linguist Ivar Aasen (1813 – 1896). Based on his studies of the Norwegian vernacular through extensive fieldwork, Aasen wrote two grammars (Aasen, 1848; Aasen, 1864) and two monolingual dictionaries (Aasen, 1850; Aasen, 1873). The 1864 grammar and the 1873 dictionary are the foundation of the written standard Nynorsk that has been one of the two official standards since 1885.

The editorial rules for documenting the vernacular in NO have their background in principles laid down by Ivar Aasen in his 1864 grammar (Aasen, 1864). He emphasized the unity and coherence of the lexicon across dialect boundaries, which primarily express themselves in phonology and morphology:

§ 375. In the lexicon, or the use of the words themselves, the landscape languages are united to a much higher degree, so that one can hardly provide a significant collection of words particular to any single small district. (Aasen, 1864, p. 356) (authors' translation)

and further:

§ 385. To the Norwegian lexicon, we count all words used in the country as long as the form of each word may be termed Norwegian, and the word form does not collide with the old rules in the language for sound positioning and word forms. (Aasen, 1864, p. 366) (authors' translation)

In short: Any dialect word form can be expressed in the standard orthography of Nynorsk, which in turn can be used to coordinate all dialect forms of a given lexical item. This also means that all NO editors must be able to evaluate and standardize Norwegian dialect materials (Gundersen, 2026, p. 83).

3.2 Editorial Assumptions and Rules

The intention behind the treatment of speech materials in NO, as expressed in the editorial rule book (Gundersen, 2016 p. 380), is

- to give speech sources the same value as written (printed) sources get,
- to give special attention to the speech materials unique to the NO language collections (contributions from informants sent in 1930 – ca. 1980),
- to source speech information explicitly when the definition or usage example in question is specific to a given part of the country,
- to indicate speech usage if a word is generally used in speech rather than print, but not specific to a given area.

Material collection for NO started in 1930 as a national joint effort. Speech materials came from voluntary informants and from researcher collections, especially from phonologists and dialectologists. The older parts of the speech collections document Norwegian vernacular from the first part of the 20th century, while newer dialect sources tend to come in print. See Section 5 below.

Is the speech information in NO valid today? The dialect map of Norway has to some extent changed. Norway underwent considerable societal changes towards the end of the 19th century. Dialect collections from before 1900 are therefore since 2002 listed separately as “older sources”, and not integrated in the (supposedly) synchronic dialect collections from after 1900 (Gundersen, 2016, p. 159). Location of usage taken from the dictionaries (Aasen, 1873; Ross, 1895) of Aasen and his closest colleague, Hans Ross, are used, but these sources are indicated by attaching an A or an R to the location name, e.g., ‘bryddaup Tel A2’ indicating that this form of *bryllaup* (‘wedding’) was located to ‘Telemark’ in Aasen (1873).

Speech sources for the central lexicon are registered in the dictionary database, but often suppressed in print. In his 1873 dictionary, which is based entirely on fieldwork, Aasen established the practice of listing the location of sources only for word forms, senses and usage examples that did not have nationwide or general coverage. Thus one finds location information only in 50% of the entries. NO continues this practice in relation to speech sources. Accordingly, entries from the central lexicon can have plentiful coverage in the speech collections, but definitions will not have location sources unless the sense itself is sparsely sourced. An example is the entry I *stein* m (‘stone noun masc.’). Out of the six senses in the entry only one, no. 6, has location sources (6. ‘floor in a fireplace; chimney base’).

What value does NO assign to speech materials in lemma selection? NO does not in principle allow entries based on a single source. The basic requirement for writing an entry is sufficient materials to identify the headword form with POS, and give it a proper sense description, in the form of a definition with a hypernym and identifying additional features. Some informants provide all that, but it is still one single source. On the other hand, the dialect materials unique to the NO language collections carry especial weight. In the end, editors are trusted to make informed judgments based on the information available to them, and NO has 55 000 single source entries – about twenty percent. Fifteen percent of these entries have a location source, signifying that the entry information is drawn from speech.

The meaning of location names in the dictionary text is simply that the information found in the Language Collections has the location listed as its place of registration. Location sourcing does not entail a claim that the word form, sense or usage is unique to the location mentioned, nor does it mean that the word is not used in print. If further documentation turns up, it will be added to the Language Collections and in time to the entry in question.

4. NO Speech Sources

The Language collections used by NO comprise:

1. Slip archives, originally on paper, digitized and available on the web. The current volume is at about 3.5 million slips. It has been estimated that the number of slips covering speech number about 0.5 million from ca. 700 informants.
2. The Dialect Synopsis – detailed form information on the pronunciation of about 1600 word forms, with the location system that NO uses.
3. The Norwegian Dialect Atlas – 596 maps showing isoglosses relevant to Norwegian dialects, created 1950–1980, digitized 2005–2007.
4. Printed dialect dictionaries. There are about 500 of them in the NO bibliography, and still more exists (Nes, 1986). Some are available only as paper books, some are also accessible via the National Library electronic bookshelf, some are transcribed and available to editors in the NO collections of electronic texts (primarily transcripts of pre 1900 texts). A remote aim is to have them all added to the Dictionary Hotel (Ordbokshotellet), a portal of dictionaries and spellers used as sources for NO.
5. The Dictionary Hotel (Ore & Grønvik, 2018) currently holds 80 dictionaries, 69 of which are dialect dictionaries with ca. 225 000 entries in all. Sixteen out of nineteen counties are represented. Northern Norway dialects have been prioritised, with 22 dictionaries in the Dictionary Hotel.
6. Speech corpora (NDK; LIA) were not available in time for the first (paper) edition of NO, which was completed in 2015. Norsk Dialektkorpus (NDK) is a dialect corpus collected 2009–2012, transcribed in a phonetic version and standardized to Bokmål (Johannessen et al., 2009, p. 74). This has caused the speech vocabulary to be misrepresented, as standard Bokmål rejects a number of much used dialect words and word forms, which entails translation to the correct Bokmål word. The speech corpus LIA Norsk v. 1.1 consists of transcriptions of dialect recordings (1939–1996) archived at Norwegian universities. LIA is standardized to Nynorsk and has in its first text version 3.5 million tokens and 1274 speakers from 226 places in Norway (Hagen & Vangsnes, 2023, p. 124). The LIA Norsk corpus was published in 2019 and is available from Språkbanken (the Language Bank) of Norway's National Library.

The collections listed above have been put together for different purposes, and are diverse in contents, organization and metadata. They support each other; the Dialect Synopsis and the Norwegian Dialect Atlas provide a general framework covering historical phonology and geographical distribution of linguistic features, while the Dictionary Hotel show a cross-country spectrum of forms and meanings indexed in standard Nynorsk form. The dialect dictionaries portray the lexicon of individual dialects. The slip archives give invaluable information, but are also the most heterogeneous in contents and metadata.

The Speech corpora NDK and LIA Norsk will be useful in documenting basic vocabulary, and especially pragmatic speech markers and the give and take of casual conversation (Askeland, 2017, p. 75), where the older collections have little to offer. The lemma inventory of the speech corpora has not yet been described in detail.

5. Quantities – Numbers of Entries and Definitions With Location of the Sources

The practice of accepting speech sources and integrating documentation from speech and writing makes NO one of a kind among the scholarly dictionaries for Nordic languages. It is therefore of interest to find out how the speech materials affect the whole.

A number of assumptions are current about the inventory of the spoken lexicon that lacks documentation in published text. The chief assumption is that speech materials deviate from the norms of the proper (standard) language, and are therefore unimportant (Sandøy, 1993, p. 11). At the same time, a number of empirically based fields of knowledge (from botany to zoology) depend on users' knowledge and terminology, whether documented in writing or not. The same applies, surely, to lexicography and linguistic fieldwork; the NO position is that documenting speech is important if the aim is to provide a valid portrait of a language through mother tongue lexicography.

The following sections will deal with how information from speech and writing is integrated at entry level, in terms of quantities and distribution, and will finally look at what sort of information the speech materials offer that is not documented in the written sources.

6. Speech Materials in Quantities and Proportions

Entries form three groups according to source type. These are firstly entries with only literary sources, secondly entries with both literary and spoken sources, thirdly entries with only spoken sources. Speech source distribution is shown in Table 2. Source type distribution is shown in Table 3.

NO entries can have information about dialect location in three parts of the entry: (1) the word form and etymology section that follows the entry head, (2) as sources for definitions, (3) as sources for usage examples. This paper focusses on source location for definitions. Within an entry, the senses are organized in a standard tree structure. The general structure is as follows: Some of the nodes (numbered senses) in the tree structure are just placeholders used to group the more finely graded senses and contain no definitions. For senses with definitions, the definitions are given as a semicolon separated list in the printed dictionary and as a list of definition fields in the database. Each definition can have one or more references to literary sources and/or references to places of the spoken sources.

As mentioned in Section 4.2, general countrywide spoken usage of a headword is not explicitly sourced. Table 2 shows that 48 percent of all entries have one or more speech sources. 51 percent of all entries have literary sources only, or lack any explicit source. Entries for the letter *c q w x z* could have been excluded, as these consonants are not

used in Norwegian orthography except in imported vocabulary, but the number of entries for these letters is so small that the outcome would not have been affected.

Of all sense sections with definitions, 46 percent have location of source. Of the single definition fields, 36 percent have location of source. The smaller percentages for sense sections and definition fields make sense; most entries with a definition with location information will often have sense sections with definitions without such information, and many sense sections have a mixture.

Table 2: Number of entries, senses and single definition fields per letter and percentage of the respective totals

Letter	Entries		Senses		Definition fields	
	Total	with location	Total	with location	Total	with location
A	5 584	38 %	8 000	36 %	10 261	28 %
B	15 643	40 %	22 859	41 %	28 860	31 %
C	204	0 %	229	0 %	270	0 %
D	8 369	44 %	14 035	44 %	17 632	33 %
E	5 945	31 %	8 549	32 %	11 576	24 %
F	27 284	32 %	42 109	30 %	54 656	23 %
G	20 751	46 %	34 611	43 %	44 982	32 %
H	19 080	49 %	31 447	47 %	42 190	34 %
I	4 599	33 %	7 490	27 %	9 186	23 %
J	3 983	46 %	5 669	43 %	6 537	37 %
K	26 157	52 %	39 959	51 %	48 505	43 %
L	15 513	48 %	22 497	47 %	28 278	39 %
M	15 157	46 %	19 709	47 %	24 686	39 %
N	7 337	48 %	10 021	48 %	12 473	39 %
O	7 045	44 %	9 854	41 %	12 594	33 %
P	10 441	40 %	13 723	43 %	17 131	35 %
Q	13	0 %	16	0 %	19	0 %
R	13 233	55 %	18 756	57 %	24 016	46 %
S	47 933	58 %	72 637	56 %	92 202	44 %
T	16 751	57 %	25 898	52 %	33 723	41 %
U	6 953	50 %	10 958	44 %	15 259	32 %
V	11 723	55 %	17 968	49 %	23 709	38 %
W	48	0 %	48	0 %	53	0 %
X	20	0 %	23	0 %	24	0 %
Y	574	59 %	916	54 %	1 341	37 %
Z	39	0 %	40	0 %	44	0 %
Æ	339	44 %	520	40 %	690	31 %
Ø	1 118	60 %	1 663	54 %	2 186	41 %
Å	1 952	62 %	2 904	57 %	3 696	45 %
Total	293 788	48 %	443 108	46 %	566 779	36 %

We have also had a look at the relative frequencies of definitions with literary sources, spoken sources and both. Table 3 gives an overview. Entries without any references to spoken or printed sources are excluded from the table. Hence the totals are smaller than in Table 2.

Table 3 deals only with entries and definition fields that are explicitly sourced. Definitions that are generally valid are not included in the total, cf. Section 4.2. Distribution quantities for written sources and speech sources in entries and definitions balance each other, with some more explicit sourcing for speech than from written sources.

Most definition fields with location sources list few locations. 93 percent list from 1 to 9 locations, 7 percent list from 10 to 99 locations, while only 107 definition fields list more than 100 locations for one definition.

Table 3: Entries and definition fields with literary and location sources, grouped by source type. Distribution shown in percent of total number.

Entries					Definition fields			
Letter	with loc. and/or lit.	with lit. only	both loc. and lit.	with loc. only	with loc. and/or lit.	with lit. only	with loc. and lit.	with loc. only
A	3 427	38 %	22 %	40 %	4 495	37 %	15 %	48 %
B	10 265	39 %	18 %	43 %	13 788	35 %	12 %	52 %
C	27	100 %	0 %	0 %	29	100 %	0 %	0 %
D	5 804	37 %	22 %	41 %	8 670	32 %	15 %	53 %
E	3 505	47 %	21 %	32 %	4 837	43 %	15 %	42 %
F	17 931	51 %	19 %	29 %	23 496	47 %	15 %	38 %
G	15 543	39 %	23 %	38 %	22 206	36 %	18 %	46 %
H	14 999	38 %	26 %	36 %	22 056	36 %	20 %	44 %
I	3 135	51 %	19 %	30 %	4 171	50 %	15 %	35 %
J	3 313	44 %	24 %	32 %	4 364	44 %	19 %	37 %
K	20 828	35 %	25 %	40 %	31 006	32 %	19 %	49 %
L	11 905	37 %	26 %	37 %	17 115	36 %	20 %	44 %
M	11 107	38 %	25 %	37 %	14 833	36 %	20 %	44 %
N	5 252	33 %	27 %	40 %	7 152	31 %	21 %	47 %
O	4 934	38 %	25 %	38 %	6 674	39 %	19 %	43 %
P	7 243	42 %	18 %	39 %	9 777	38 %	14 %	48 %
Q	3	100 %	0 %	0 %	3	100 %	0 %	0 %
R	1 0103	28 %	27 %	45 %	14 604	25 %	21 %	54 %
S	37 069	25 %	27 %	47 %	53 702	24 %	21 %	56 %
T	13 097	27 %	27 %	46 %	18 298	25 %	21 %	54 %
U	5 017	30 %	28 %	42 %	6 808	29 %	23 %	48 %
V	9 025	28 %	29 %	43 %	12 217	27 %	23 %	50 %
W	22	100 %	0 %	0 %	23	100 %	0 %	0 %

X	11	100 %	0 %	0 %	11	100 %	0 %	0 %
Y	507	33 %	34 %	34 %	710	30 %	27 %	42 %
Z	18	100 %	0 %	0 %	20	100 %	0 %	0 %
Æ	274	45 %	30 %	25 %	371	42 %	23 %	35 %
Ø	903	26 %	29 %	45 %	1 211	25 %	24 %	51 %
Å	1 687	28 %	30 %	42 %	2 267	27 %	23 %	50 %
Total	216 954	35 %	25 %	40 %	304 914	33 %	19 %	48 %

7. Words and Senses From the Vernacular – a Sample

The number of definition fields with speech sources only is close to 150 000. As a sample, we have selected the definition fields sourced from the same locations as the 69 dialect dictionaries in the Dictionary Hotel, with 12 595 lines drawn from 11 183 different entries.

In the sample of definitions, grammatical markers and other abbreviations have been removed to facilitate sorting by first word in the definition proper. Headword form with POS will show the lexicon profile of the sample. A set of definitions will also reflect the definition format register for the dictionary in question, which in turn indicates type of word and sense.

Speech is assumed to differ from writing in using more verbs and in using fewer compounds, especially compound nouns. If this assumption is correct, definitions sourced from speech materials should be found under a group of headwords reflecting their speech origins.

The POS distribution in the sample is 17 % definitions for adjectives and adverbs, 64 % nouns and 15 % verbs. In NO as a whole, the distribution is different: the dictionary has 15 % adjectives and adverbs, 76 % nouns and 7 % verbs. In the two standard dictionaries for Bokmål and Nynorsk, entries for verbs are between 9 and 10 % of the total. The almost double number of verbs in the headword list of the sample is therefore significant.

NO distinguishes between compounds on one side and simple and derived word forms on the other. In NO 75 % of headwords are compounds, 25 % simple or derived headword forms. NO has 9 600 entries for multi-word expressions (MWE), corresponding to about 4 % of the total number. Word type distribution in the sample headwords is 22 % simple word forms, 25 % derived word forms, 49 % compound word forms, and 4 % MWE. The significant difference lies in the proportion of compounds: 75 % in NO as a whole, 49 % in the sample of definitions with speech sources only. It seems that simple or derived word forms play a much larger part in speech than in writing. The headwords of the sample definitions strengthen the general view of word types in speech versus writing.

Headword selection by informants and editors of dialect dictionaries tends to be weighted towards content POS. Function words (prepositions, conjunctions,

subjunctions, determinatives etc.) are almost the same in Bokmål and Nynorsk, and common to all speakers of Norwegian. Function words are rarely mentioned in lexical documentation of vernaculars, while content words (nouns, adjectives and adverbs, verbs, MWEs) dominate. The tendency in the sample analysed here corresponds with what is thought to be important differences between the lexicon of speech and the lexicon of writing.

Another assumption concerns word origins, or etymology. Written language is taken to have more lexical items imported from other languages (“hard words”), whereas the speech lexicon is supposed to stay closer to the basic lexicon, learnt early and known to everyone in a language community. In Norway, imported lexical items have arrived from a number of languages, and the Greek and Latin influence via English (and formerly French) is strong in written language and the media. In contrast, there are few imported lexical items in the sample, either in headword simple forms or in derivation endings. Simple word forms are monosyllabic or disyllabic and have the phonological structure expected for Norwegian base forms. A majority of adjectives are derived; the dominant endings are *-al*, *-en*, *-ig*, *-leg*, *-sam*, *-ut*, all of which belong to the Norwegian standard repertory.

The formats of full definitions correspond to the POS of the headword. The sample shows no different use of definition formats from entries with mainly literary sources.

An NO definition can start with an extension in brackets and a grammatical marker before the definition itself. A full definition has a hypernym with additional distinguishing features. NO also uses synonym definitions where several words have an identical meaning. The less used word forms are defined with the more frequent word, which has a full definition. The total of synonym definitions in NO is 48 643, about 16 per cent of all sourced definitions fields. In the sample 4250 definitions are synonym definitions, 33,7 per cent, or a third. This is reasonable; the sample is drawn from a limited register of speech sources. It is noteworthy that the sample also holds the main definitions for 12 plant species. This shows that the NO written sources do not document all standard plant names for species – yet.

A full definition can be seen as a status marker for the headword in question; if the headword in a given sense has a full definition and also full synonyms, it is the most important word form in its peer group. A species name used in schoolbooks will have a full definition; less used names for the same species are often synonym defined, i.e., with a single word. The standard Norwegian word for the plant dandelion, “løvetann” is linked to 114 entries through synonym definitions. The web version of NO shows the list of synonyms under the heading “Tilvist frå” (‘cross referenced from’).

8. Meanings and Fields of Knowledge

The meanings and fields of knowledge reflected in the sample are too diverse for comparing proportions at the present stage, but some observations can be made.

The headwords in the sample refer to concrete objects rather than abstracts. The evidence comes from definition extensions and the use of adjectives qualifying hypernyms.

The initial extension in brackets uses the formula “(om ...)” ‘concerning ...’. It is found in ca. 200 sample definitions. In the extension, very few abstract nouns are found in what follows “om”. In NO as a whole, one finds f. i. “om framferd, handling, kjensle, mengde, regel” (‘concerning behaviour, action, emotion, quantity, rule’). Hypernyms for classes of concepts are also frequent, f. i. “om dyr, landskap, reiskap” (‘concerning animals, landscape, tools’). These are not found in the sample, with one exception “om arbeid” ‘about work’.

A large number of words refer to objects distinguished by their size, shape or any other measurable quality. Definitions starting with the adjective opposites *stor* – *liten*, *små* (‘large – small’), *lang* – *kort*, *stutt* (‘long – short’), *tunn* – *tjukk* (‘thin – thick’), *tung* – *lett* (‘heavy – light’) are frequent. The definitions refer to the physical qualities of objects or living beings.

Some topic areas come forward as important in the sample, that are less visible in NO as a whole. Such topics are landscape, its shape and usefulness; wood and stone as (pieces of) objects and material; and equipment for sorting, treating and transporting natural produce.

These sense sections can be very specific, and their communicative value requires context. An example:

«**steinslag** stein som har slegi mot bartre og grodd inn i treet»

‘**stoneblow** stone that has hit a pine tree and grown into the tree’

This is an important combination of headword and definition if you plan to fell trees in a steep and scree-exposed landscape, as felling a tree with «steinslag» could start a scree. Norway is a country where life is shaped by topography, and wood, especially timber, has been a major source of activity and income at least since the early Middle Ages. But there is no Norwegian terminological dictionary for wood and woodworking, except for an addendum in a book on log building (Strømshaug, 1997). At a guess, the terminology of wood is integrated into the Norwegian language to the point where it is not thought of as a separate field of expertise. The alternative possibility is negligence and omission.

Many definitions refer to persons and animals, and characterize appearance, behaviour and habits. The main impression from these definitions is that they represent assessment of qualities or usefulness rather than personal and moral judgments. Negative judgments can shine through, partly in defining language and partly in headword form. If a compound ends in *-fant* (masculine) ‘(young) man (without a settled home)’, the definition is unlikely to be complimentary. The sample, comprising 11 000 entries, has 38 headwords with the element *fant*; NO as a whole has 471 out

of around 300 000 headwords, suggesting that this characterisation is more used in speech than in writing.

9. Conclusion

This article gives an overview of the way spoken source material is used and documented in *Norsk Ordbok* (NO). We also looked at the entries of NO via the source system in order to see if there are visible and measurable marks of origin separating information from speech sources from information from written sources. Significant differences are found in relation to word typology, distribution of POS, the use of definition formats and subject matter.

This paper is a preliminary investigation. Results suggest that more can be found, both in content and in method development. The major dictionaries with their wealth of linked headwords and definitions have a part to play in ensuring natural language a place in a digital future where both speech and writing need correct interpretation and use.

References

- Aasen, I. (1850). *Ordbog over det Norske Folkesprog* (Dictionary of the Norwegian Vernacular).
- Aasen, I. (1864). *Norsk Grammatik* (Norwegian Grammar). P. T. Mallings Forlagsboghandel.
- Aasen, I. (1873). *Norsk Ordbog med dansk forklaring*. (Norwegian Dictionary with Explanations in Danish) Christiania. Mallings Boghandel.
- About the Norwegian Dialect Synopsis* (Målføresynopsisen). Retrieved May 27, 2024, from https://www.edd.uio.no/synops/work/synops_ext/Opplysningstekstar_eng.html
- Askeland, A. (2020). *Serdu: eit attersyn og eit nytt perspektiv* (*You see: a Review and a New Perspective*). In O. Almenningen, O. Grønvik, L. S. Vikør, & D. Worren (Eds.), *Leksikografen frå Leksvika* (pp. 75–82). Novus forlag.
- Dictionary of Finnish dialects*. Retrieved January 30, 2024, from https://www.kotus.fi/en/dictionaries/dictionary_of_finnish_dialects
- Grønvik, O. (2015). Language Documentation in a Globalised World. In J. Ragnar Hagland, & Å. Wetås (Eds.), *Ivar Aasen ute og heime – om moderne språkdokumentasjon etter Ivar Aasen*. (pp. 28–53). Skrifter 2015 nr 1.
- Gundersen, H. (2016). *Redigeringshandbok for Norsk Ordbok 2014* (Editorial Guidelines for Norsk Ordbok 2014). Retrieved May 27, 2024, from <http://no2014.uib.no/eNo/tekst/redigeringshandboka/redigeringshandboka.pdf>
- Hagen, K., & Vangsnes, Ø. A. (2023). LIA-korpora – eldre talemålsopptak for norsk og samisk gjort tilgjengelige (The LIA Corpora – Older speech recordings of Norwegian and Sami made accessible). *Nordlyd* Vol. 47 No. 2 (2023). *Struktur, ideologi og mangfald*. <https://septentrio.uit.no/index.php/nordlyd/article/view/7157/7836>

Hellevik, A. (1956). *På skattegraving i eige mål* (Treasure Hunting in one's own Language). Syn & Segn. 1956 Vol. 62.

Johannessen, J. B., Priestley, J., Hagen, K., Åfarli, T. A., & Vangsnes, Ø. A. (2009). The Nordic Dialect Corpus – an Advanced Research Tool. In K. Jokinen, & B. Eckhard (Eds.), *Proceedings of the 17th Nordic Conference of Computational Linguistics NODALIDA 2009*. NEALT Proceedings Series Volume 4. <https://aclanthology.org/W09-4612/>

LIA Norsk v. 1.1. (Norwegian dialect corpus v. 1.1). Retrieved May 27, 2024, from <https://www.nb.no/sprakbanken/en/resource-catalogue/oai-tekstlab-uio-no-lia-norsk/>

Målføresynopsisen (Dialect Synopsis for the Norwegian Dialects). Retrieved May 27, 2024, from <https://usd.uib.no/perl/search/search.cgi?appid=145&tabid=2165>

Nes, O. (1986). *Norsk dialektbibliografi* (Norwegian Dialect Bibliography). Novus Forlag.

Netland, R., & Opsahl, T. (2024). Dialekt in *Store Norske Leksikon på snl.no*. Retrieved May 27, 2024, from <https://snl.no/dialekt>

Nordisk Dialektkorpus (NDK) (Nordic Dialect Corpus). Retrieved May 27, 2024, from <https://tekstlab.uio.no/nota/scandiasyn/>

Norsk dialektatlas (Norwegian Dialect Atlas) Kartene. Retrieved May 27, 2024, from <https://usd.uib.no/perl/search/search.cgi?appid=210&tabid=2338>

Norsk Ordbok. Ordbok over det norske folkemålet og det nynorske skriftmålet (*The Norwegian Dictionary*. Dictionary of the Norwegian vernacular and the Nynorsk written standard). 1966–2016. Vol. 1–12. Oslo. Det Norske Samlaget.

Norsk Ordbok. Retrieved May 27, 2024, from <https://alfa.norsk-ordbok.no/>

Norsk Ordbok AH 2005 (<https://usd.uib.no/perl/search/search.cgi?appid=251&tabid=3581>)

NRK-språket. Språkreglar (nynorsk) 2007. (National Broadcasting Language. Guidelines (Nynorsk)). Retrieved May 27, 2024, <https://sprak.nrk.no/rettleteing/nynorsk/>

Ordbokshotellet (The Dictionary Hotel). Retrieved May 27, 2024, from <https://usd.uib.no/perl/search/search.cgi?appid=118&tabid=1777>

Ore, C.-E. S., & Grønvik, O. (2018). Comparing Orthographies in Space and Time through Lexicographic Resources. In *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts* (pp. 159–172). Znanstvena založba Filozofske fakultete.

Ross, H. (1895). *Norsk Ordbog*. Tillæg til “Norsk Ordbog” af Ivar Aasen (Norwegian Dictionary. Supplement to “Norsk Ordbog” by Ivar Aasen). Kristiania.

Sandøy, H. (1985). *Norsk dialektkunnskap* (Norwegian Dialectology). Novus Forlag.

Sandøy, H. (1993). *Talemål* (Spoken Language). Novus Forlag.

Setelarkivet til Norsk Ordbok (The slip archive of Norsk Ordbok). Retrieved May 27, 2024, from <https://usd.uib.no/perl/search/search.cgi?appid=8&tabid=436>

Skard, S. (1932). Norsk Ordbok. Historie – plan – arbeidsskipnad. (Norsk Ordbok. History – Plan – Organisation). In H. U. Karlsen, L. S. Vikør, & Å. Wetås (Eds.), *Livet er æve, og evig er ordet: Norsk ordbok 1930–2016* (pp. 40–67). Det Norske Samlaget.

Strømshaug, K. (1997). *Lafting. Emne og omgangsmåte* (Log Building. Materials and Handling). Landbruksforlaget.

Tier, V.d., & Keymeulen, J.v. (2010). Towards the completion of the Dictionary of the Flemish Dialects. Ghent University (Belgium). In A. Dykstra, & T. Schoonheim (Eds.), *Proceedings of the XIV EURALEX International Congress. 6–10 July 2010*. Fryske Akademy – Afûk.

Wright, J. H. (Ed.) (1981). The English Dialect Dictionary. Vol. I–VI. Oxford University Press. Retrieved January 20, 2024, from <https://www.uibk.ac.at/en/disc/blog/english-dialect-dictionary-online/>

Aasen, I. (1848). *Det Norske Folkesprogs Grammatik* (Grammar of the Norwegian Vernacular). Det Kongelige Norske Videnskabers Selskab. Werner & Comp.

Contact information

Christian-Emil Smith Ore

University of Oslo, Dept. of Linguistics and Scandinavian Studies
c.e.s.ore@iln.uio.no

Oddrun Grønvik

University of Bergen, Department of Linguistic,
Literary and Aesthetic Studies
Oddrun.Gronvik@uib.no

Petya Osenova and Kiril Simov

ALL ABOUT WORDS! AN INTEGRATED DICTIONARIES PORTAL FOR BULGARIAN

Abstract The paper introduces a web portal of integrated dictionaries for Bulgarian. The mapping among the resources is lemma-based. Two dictionaries are in the centre of this integration – an Inflectional dictionary of Bulgarian, since Bulgarian is a morphologically rich language, and a Wordnet of Bulgarian – BTB-Wordnet, since it adds the level of lexical meanings to the dictionary-enhanced knowledge. Also, various other types of dictionaries are being gradually added – with diachronic spellings, bilingual, specialised, etc.

Keywords Bulgarian; dictionaries; integration

1. Introduction

The tradition in dictionary making and consequently – dictionary access, favoured the following rule: for each language phenomenon – a separate dictionary, mainly on paper. Thus, there were dictionaries for spelling, synonyms, antonyms, derivational relations, explanatory dictionaries with senses, new words and senses, jargon, dialectal, of foreign words, etc. Then the time of thesauri and linked dictionaries came – wordnets and ontowordnets were created for many languages; also babelnets, verbnets, framenets, wikidata, wiktionaries, etc. However, very often the users cannot get oriented where to find the most appropriate dictionaries for their specific tasks, especially when they need to consult not only contemporary dictionaries, but also historical ones, or when they need more complex information accessible in more than one dictionaries. Within CLaDA-BG infrastructure¹, we have to provide access to dictionaries that are in active use by different researchers in their everyday work. For example, ethnographers and historians very often explore documents written in different time periods and in different language varieties of Bulgarian. Thus, they consult dictionaries published more than one hundred years ago. In order to provide access to such dictionaries one of the tasks is to scan, OCR and proofread old dictionaries of Bulgarian, published within the period from the second half of the nineteenth century to the middle of the twentieth century. In order to facilitate the access to these dictionaries, our solution was to construct their online representation and to connect them to the contemporary lexical resources.

Thus, we decided to provide a web-based portal for users in Bulgarian which integrates a number of dictionaries with interlinked content.² The portal is based on a vocabulary of lemmas shared among dictionaries. For each lemma its paradigm (if available) was stored and linked to the corresponding lexical entries

¹ <https://clada-bg.eu/en/>

² <https://concordance.webclark.org/#/word/гледях>

in the other dictionaries. The user can navigate within the vocabulary through: i) access to a list of lemmas starting with a specific prefix; ii) a specific lemma; iii) a word form which first undergoes lemmatization and then is searched in the vocabulary. Here a central role plays the *Inflectional dictionary of Bulgarian* (IDB).³ It provides information not only about the lemma (with its part-of-speech), but also about its paradigm and grammatical features. The next important dictionary-thesaurus is *BTB-Wordnet*.⁴ We put this resource in the center of our dictionary hub because it covers a carefully selected set of senses and thus provides a basis for clustering information on sense level from the other dictionaries within the system. In the current version of the portal the alignment has been performed on the basis of lemmas, but in future we consider also an alignment on sense level. Although the lemma level seems to be an easy mechanism for alignments, there are two main problems to be solved – at least in our case. The first problem is that some of the dictionaries are results from OCR-ed paper dictionaries or based on paper dictionaries, and others are already digitally born. Thus, the approaches to lemmatization very often differ – especially with respect to inclusion or exclusion of optional reflexive particles or to considering attributively used participles as a separate part-of-speech. The second problem is the dynamics in the development of the Bulgarian language spelling system in the period from the second half of the nineteenth century to the middle of the twentieth century. This fact caused some difficulties in word normalisation and the consequent lemmatization. Thus, intensive work was invested from our side to unify the lemma representation across all the available resources.

The structure of the paper is as follows: the next section outlines some conceptual foci of the related work. Section 3 gives an overview of the integrated dictionaries model. Section 4 outlines the implementation details of the web portal. Section 5 concludes the paper.

2. Related Work

The idea of integrating and interlinking various grammatical and lexical resources has been explored successfully for quite some time now for various languages and various tasks. Here we mention only a few of the related works to show what kind of approach we follow at the moment, and what our aims are in the near future.

Herold et al. (2012) introduce a complex typed link structure to ensure the proper relations among large modern and historical dictionaries with differing lexicographic traditions. The mapping has been performed through an aligned lemma list where the strategies for achieving entry equivalence are presented. Our approach is similar and we ensure the mappings through a lemma list, unified across resources.

³ The link is with an example: the paradigm of the word form *гледax* (gledax-1PERS.SG.AORIST/IMPERFECT, I looked/was looking): <https://concordance.webclark.org/#/word/%D0%B3%D0%BB%D0%B5%D0%B4%D0%B0%D1%85>

⁴ The link is with an example – the word *глад* (glad, ‘hunger’): <https://concordance.webclark.org/#/wordnet/глад?synID=14281&pos=n>

In Tiberius et al. (2021, p. 60), the ELEXIS dictionary matrix has been discussed among other initiatives. It is defined as “universal repository of linked senses, meaning descriptions, collocations, phraseology, translation equivalents, examples of usage and other types of lexical information found in existing lexicographic resources, monolingual, multilingual, modern, historical etc.” In addition, the constituting elements of a common vocabulary have been presented. Our model is not sense-based yet but it comprises most of the characteristics of the common vocabulary such as an entry, a headword, a part-of-speech, a sense, a sense structure, a definition, an example, information about the inflected forms. More details of our model are given in the next section.

In Ahmadi et al. (2020, p. 3234), a sense-alignment strategy is presented for 15 languages, among which Bulgarian. The lemmas were selected through specially designed criteria. The aim was “to provide semantic relationships between two sets of senses for the same lemmas in two monolingual dictionaries.” For Bulgarian these were the BTB-Wordnet resource and the Wiktionary. In our integrated portal the mapping has not been performed through semantic relations yet but at the same time it refers to semantic resources and thus is semantics-aware.

3. Overview of the Integrated Dictionaries Model

The main components of the system are as follows: *Words in context* (this is the related concordancing service where one can search for word examples) and *All about words* (where information is provided about the grammatical features of words; about the synsets, they are part of; about their definitions, links to various dictionaries, etc.). If the user starts with the *All about words* component, they can write the word of interest in the search box. Even when it is not a lemma but a word form, the result will contain information about its lemma(s) as well. The integrated structure includes the following information: a) a part-of-speech; b) a grammatical paradigm, if any; c) a grammatical description; d) all the senses available in the wordnet; e) examples from the included corpora; f) relation to the existing etymological dictionary; g) relation to the existing terminological dictionaries, if applicable; h) relation to a ‘diachronized’ version of the word, if different from the contemporary spelling. Let us illustrate our integrated dictionary access with the word дете (dete, child, ‘a child’). From the large inflectional dictionary, the following information displays: де|те, ~тето, ~ца, ~цата. The vertical marker points the place in the lemma where it changes when it gets its word forms: a definite singular form (child-the ‘the child’), a non-definite plural form (children) and a definite plural form (children-the, ‘the children’). Then a piece of grammatical description is available from this dictionary about the form that has been selected by the user. In this case the description says: съществително нарицателно, среден род, единствено число, нечленувано ‘noun, common, neuter, singular, non-definite’. The BTB-Wordnet provides 4 meanings and 9 examples. At the moment the examples are listed separately from the meanings in the user interface, but it is easy to show the links between the examples and the meanings.

The lexical entries of the integrated lexicons have the following structure: (1) *Inflectional dictionary of Bulgarian*. The lexical entry in this lexicon consists of the following elements: < Lemma, POS, Paradigm, Inflectional class, Clitics >, where the *Lemma* coincides with the basic form of the word; *POS* is the part-of-speech of the lemma; *Paradigm* is a list of all word forms for the lemma where each word form is associated with a tag, representing its grammatical features; *Inflectional class* is an index to all lemmas that share the same inflectional rules; *Clitics* provides a list of obligatory or optional reflexive clitics for the lemma (usually for the verb lemmas). Currently, IDB contains around 81 238 lemmas with 1 485 026 word forms. We use an XML structure for the representation of the lexical entries. (2) *BTB-Wordnet* (Simov & Osenova, 2023) is a Wordnet for Bulgarian encoded in WN-LMF⁵. BTB-Wordnet contains 35 833 synsets and more than 50 000 lemmas. BTB-Wordnet is aligned to the Open English Wordnet.⁶ Also, the lemmas in BTB-Wordnet are aligned to the lemmas in IDB. Each lemma within a given synset is related to examples that exemplify the corresponding sense. Currently, we have gathered nearly 80 000 examples. Our goal is to assign at least 3 examples per lemma within the synsets. (3) *The Bulgarian part of the PONS bilingual dictionaries*. This part was created in-house and the publisher left the copyright of the Bulgarian vocabulary to us. In this case the lexical entries have the following structure: < HeadWord, POS, Grammatical information, Inflectional Representation, Senses, Phraseology >. The lexical entry format is inspired by the Concede lexical model.⁷ This dictionary contains about 45 000 lexical entries and covers more than 60 000 senses. For IDB and BTB-Wordnet (the core lexicons of the integrated lexicons portal) we map the head words with the lemmas in IDB and BTB-Wordnet. Here this mapping was done already during the creation of the Bulgarian vocabulary.

Thus, currently we use the lemmas from each added dictionary as a means for the integration. It should be noted that some lemmas are represented in just one of the dictionaries. In order not to lose data we maintain a joint list of lemmas covering all the dictionaries in the portal. When the user searches for a certain lemma, only the information from the corresponding dictionaries is displayed in the user interface. The lemmas in each dictionary are used for keys which are stored in the joint database of all dictionaries.

Similarly, we define an appropriate XML representation for the elements of the lexical entries in the corresponding dictionary. In this way, we have a database representation of the dictionary which allows us to check and edit the dictionary content. Thus, for each dictionary, we first construct an appropriate representation in XML, through which the appropriate structure of the lexical entries is reflected. After having represented the lexical entries of some dictionaries, we aim at unifying their lemmas. In order to maintain digitised dictionaries published before 1945 and within the old Bulgarian orthography, we also created a list of lemmas according to this orthography with corresponding lemmas in the contemporary vocabulary.

⁵ <https://github.com/globalwordnet/schemas>

⁶ <https://en-word.net/>

⁷ https://www.researchgate.net/publication/2403685_The_Concede_model_for_lexical_databases

Usually, the mapping is done by looking up in the list of old lemmas as a first step. If the candidate lemmas are not presented within the list, we add them to it, convert them to the new spelling (via a set of rules) and a specialist checks the result.

In the extension phase of the web portal, we envisage that each dictionary has its own webpage and thus might be consulted separately from the rest of the portal. At the moment, such a separate webpage is implemented for the BTB-Wordnet only. The switch to the webpage is available from the list of meanings presented within the *All about words* component.

Two other contemporary dictionaries for Bulgarian to be included in the near future are: the Valency dictionary (Osenova, Simov, Laskova & Kancheva, 2012) and the Hurltex dictionary. At the moment their verb vocabulary is being mapped to the wordnet lemmas and meanings, and the resources are being further edited and cleaned. The Valency dictionary represents the combinatorial potential of the Bulgarian verbs. It is also related to the BTB-Wordnet and to the English Verbnet. The Hurltex dictionary started from the multilingual one presented in Bassignana, Basile & Patti (2018). The Bulgarian part was downloaded and cleaned. Then each lemma was connected to the appropriate sense in BTB-Wordnet. When missing there, the specialists added these senses themselves.

Besides these dictionaries, currently we have scanned, OCR-ed and converted to XML presentation some historical dictionaries. These dictionaries are also aligned to the joint list of lemmas. The dictionaries are: (1) *Najden Gerov. Dictionary of the Bulgarian Language with interpretation of speeches in Bulgarian and Russian, volumes 1 to 5 (published at the end of 19th century and the beginning of 20th century)*. It is an explanatory dictionary, based on materials collected by the author himself for a period of 50 years. Besides the definitions, the dictionary contains information about the usages of the words as well as related proverbs, songs and others. (2) *Stoyan Romanski. Spelling dictionary of the Bulgarian literary language. 1933*. (3) *Stefan Mladenov. Etymological and orthographic dictionary of the Bulgarian literary language. 1941*. All these dictionaries are written in the old Bulgarian spelling system. In order to support the mapping between the lemmas in these dictionaries and the contemporary ones we maintain a representation of the joint lemma list in both spellings of Bulgarian. The joint list of lemmas is based on the *Inflectional lexicon of Bulgarian*. Thus, it contains the whole paradigms of most of the lemmas. The representation of the list in the old spelling system is presented (as mentioned above) in a separate dictionary of Bulgarian.

The provided combined functionality is somewhat similar to the word sketch oriented approach as defined by (Kilgarrieff & Tugwell, 2001). However, the resource described here is not meant for lexicographers only, but also for any interested parties – specialists in Digital Humanities, students, interested parties, etc. We view the word sketch through its contexts of usage as well as through its morphosyntactic profile, its meanings, its diachronic stages, its terminological load, its register, domain, etc.

4. Web Portal Implementation Outline

There exists a CLaDA-BG dictionary system (Angelov, Simov, Osenova & Kancheva, 2022) which is used for dictionary creation and editing. It also includes some of the dictionaries present in the web portal of integrated dictionaries (such as BTB-Wordnet, Explanatory Dictionary of Bulgarian, Inflectional dictionary, and some others). The resources in the web portal can be used for easy searches and consultations. The portal displays resources we have the publishing rights for. One drawback is that at the moment the two services – the editing and the web ones – are separately implemented. Thus, as future work we plan to unify them. The first step is the design and implementation of a shared database. Currently, for the server we are using MariaDB⁸, Spring Boot framework⁹ and Java 17.

As it was described above, for each entry in a given dictionary one or more keys are constructed. The keys are matched to the joint list of lemmas of the inflectional dictionary and the wordnet. This list is extended with new lemmas when they are not present in the two central dictionaries. In this way all the dictionaries are arranged around this joint list of lemmas. If the key has no lemma to correspond to, then it is inspected by a specialist and added to the joint list of lemmas. It is also marked for future analysis. Such lemmas could be of two main sorts: (1) a contemporary lemma for words that are not presented within both - the inflectional dictionary and the wordnet; and (2) it is a lemma of an old word, which is not actively used in contemporary Bulgarian due to some old spelling, or of a word which is not used at all and thus not recognised by most Bulgarians at the moment. Such lemmas are marked as archaic.

After the assignment of keys for each lexical entry in a given dictionary, the data is loaded into a relational database which includes an index on the lemmas. For each lexical entry the following information is stored in the databases: (1) a visual representation of the lexical entry; (2) examples related to the lexical entry; (3) the dictionary from which the lexical entry originates. For the user interface we rely on Bootstrap for UI¹⁰ and plain Javascript.

⁸ <https://mariadb.org/>

⁹ <https://spring.io/projects/spring-boot/>

¹⁰ <https://getbootstrap.com/>

Figure 1 presents the user interface to the *All about words* module.

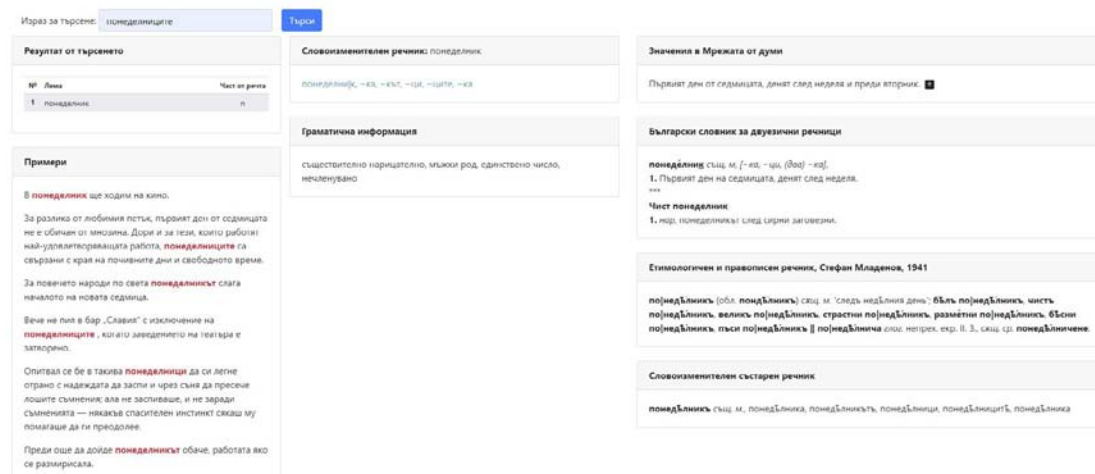


Fig. 1: The user interface to the *All about words* module

The input for the module is a search string. It can be any Cyrillic string. The search is processed in the following way: (1) first the string is analysed by the API for a word forms analysis. The API returns one or more lemmas for the word form. In this case, the lemmas are presented in the list of lemmas on the window in an alphabetical order; (2) if the string is not a (known) word form and cannot be analysed by the API, the system checks whether it is a prefix of a lemma. If the string is a prefix of such a lemma (or more than one), then the lemmas that start with the string are listed in the lemma box on the user interface. (3) If the string is not a word form or a prefix, then a message is displayed that there is no result. When the list of lemmas is displayed, the information related to the first lemma is presented. In the middle of the screen, the paradigm of the lemma is given (in case the lemma is included in the inflectional dictionary). Here the paradigm is given for the noun *понеделник* (*ponedelnik*, 'Monday'). Below the paradigm information, the user might consult the grammatical features associated with each of the word forms. On the right side of the screen, the lexical entries from the different dictionaries are presented. First, the synsets from the wordnet are shown (in Figure 1 just one synset is presented, followed by the icon which allows for switching to the website of the wordnet). There more information about the synset is outlined. Then the lexical entries from the other dictionaries are shown. The last one in the list is from the etymological dictionary and it demonstrates the lexical entry in the old Bulgarian spelling (from before 1945). On the left of the screen, under the list of lemmas, a list of the available examples is given. The examples in this case are the ones related to the synsets in the wordnet as it was mentioned above.

5. Conclusions

In this paper, we presented a lemma-based model for integrating various dictionaries for Bulgarian in an online portal. We believe that this portal answers the needs of the users about getting knowledge from multiple dictionaries and within the context of other available lexical resources.

The lemma-based approach that we follow here is far from straightforward, especially in cases where diachronic data have been added. While the lemma-oriented mapping is very important for dictionary integration, it is not sufficient as a step. For that reason, as future work we plan to exploit some mixed models of dictionary mapping, i.e., lemma- and sense-based, example-based, user-oriented, etc.

References

- Ahmadi, S., McCrae, J. Ph., Nimb, S., Khan, F., Monachini, M., Pedersen, B., Declerck, Th., Wissik, T., Bellandi, A., Pisani, I., Troelsgård, Th., Olsen, S., Krek, S., Lipp, V., Váradi, T., Simon, L., Gyorffy, A., Tiberius, C., Schoonheim, T., Ben Moshe, Y., Rudich, M., Abu Ahmad, R., Lonke, D., Kovalenko, K., Langemets, M., Kallas, J., Dereza, O., Fransen, Th., Cillessen, D., Lindemann, D., Alonso, M., Salgado, A., Sancho, J. L., Ureña-Ruiz, R.-J. Zamorano, J. P., Simov, K., Osenova, P., Kancheva, Z., Radev, I., Stanković, R., Perdih, A., & Gabrovsek, D. (2020). A Multilingual Evaluation Dataset for Monolingual Word Sense Alignment. In N. Calzolari, F. Béchet, Ph. Blache, Kh. Choukri, Ch. Cieri, Th. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 3232–3242). European Language Resources Association.
- Angelov, Zh., Simov, K., Osenova, P., & Kancheva, Z. (2022). The CLaDA-BG Dictionary Creation System: Specifics and Perspectives. In T. Erjavec, & M. Eskevich (Eds.), *CLARIN Annual Conference Proceedings* (pp. 24–28).
- Bassignana, E., Basile, V., & Patti, V. (2018). Hurtlex: A Multilingual Lexicon of Words to Hurt. In El. Cabrio, Al. Mazzei & F. Tamburini (Eds.), *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)* (pp. 51–56). Accademia University Press.
- Herold, A., Lemnitzer, L., & Geyken, A. (2012). Integrating Lexical Resources Through an Aligned Lemma List. In C. Chiarcos, S. Nordhoff, & S. Hellmann (Eds.), *Linked Data in Linguistics* (pp. 35–44). Springer.
- Kilgariff, A., & Tugwell, D. (2001). WORD SKETCH: Extraction and Display of Significant Collocations for Lexicography. In *Proceedings of the ACL Workshop on COLLOCATION: Computational Extraction, Analysis and Exploitation* (pp. 32–38). Association for Computational Linguistics.
- Osenova, P., Simov, K., Laskova, L., & Kancheva, S. (2012). A Treebank-driven Creation of an OntoValence Verb lexicon for Bulgarian. In N. Calzolari, Kh. Choukri, Th. Declerck, M. Uğur Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation* (pp. 2636–2640). European Language Resources Association (ELRA).
- Simov, K., & Osenova, P. (2023). Recent Developments in BTB-WordNet. In G. Rigau, F. Bond, & A. Rademaker (Eds.), *Proceedings of the 12th Global Wordnet Conference* (pp. 220–227). Global Wordnet Association.
- Tiberius, C., Krek, S., Depuydt, K., Gantar, P., Kallas, J., Kosem, I., & Rundell, M. (2021). Towards the ELEXIS data model: defining a common vocabulary for lexicographic resources. In I. Kosem, M. Cukr, M. Jakubíček, J. Kallas, S. Krek, & C. Tiberius (Eds.), *Electronic lexicography in the 21st century. Proceedings of the eLex 2021 conference* (pp. 56–77). Lexical Computing CZ.

Acknowledgements

The reported work has been partially supported by CLaDA-BG, the Bulgarian National Interdisciplinary Research e-Infrastructure for Resources and Technologies in favor of the Bulgarian Language and Cultural Heritage, part of the EU infrastructures CLARIN and DARIAH, funded by the Ministry of Education and Science of Bulgaria (support for the Bulgarian National Roadmap for Research Infrastructure). We would like to thank also the reviewers for their valuable and insightful comments.

Contact information

Petya Osenova

Institute of Information and Communication Technologies
Bulgarian Academy of Sciences
petya@bultreebank.org

Kiril Simov

Institute of Information and Communication Technologies
Bulgarian Academy of Sciences, Bulgaria
kivs@bultreebank.org

Vanja Štefanec, Krešimir Šojat, and Matea Filko

CRODERIV – SEARCH AND VISUALIZATION INTERFACE

Abstract In this paper, we present the search and visualization interface of the Croatian derivational lexicon – CroDeriv. CroDeriv contains information on the derivational and morphological properties of Croatian lexemes. Each lemma in the lexicon is enriched with its word-formation analysis and morphological segmentation. The search interface enables simple and advanced queries, i.e., by lexemes, by morphological structures, and by word-formation patterns. Moreover, the visualization interface enables the graphical representation of derivational families.

Keywords CroDeriv; Croatian language; morphological analysis; word-formation; morphemes; derivational family

1. Introduction

In this paper, we present the search and visualization interface for CroDeriv¹, an online lexicon that contains data on the morphological structure and derivational relatedness of Croatian lexemes.

Although there are numerous language resources dealing with inflectional morphology for various languages, including Croatian (e.g., Croatian Morphological Lexicon (Tadić & Fulgosi, 2003); hrLex (Ljubešić et al., 2016)), language resources dealing with word-formation morphology are very scarce (see Kyjánek, 2018; Filko, 2020, pp. 109–117). The main difference between derivational resources developed for other languages and CroDeriv is that the former focus almost exclusively on the derivational relatedness of lexemes, disregarding the description of their morphological structure. CroDeriv is therefore a unique language resource on the morphological level since it comprises data on word formation as well as morphological structure. Each lexeme in CroDeriv is segmented into the morphemes it consists of, and each lexical entry also contains information about the derivational relatedness between itself and its derivational base. This kind of representation enables various search (and research) possibilities, which would not be feasible with other derivational resources mentioned above. To make various complex queries across the lexicon possible for a broader audience of users, we developed an intuitive and user-friendly search and visualization interface.

In the following section, we briefly present CroDeriv and the principles of its organization. In Section 3, we describe the search and visualization interface. The paper ends with the concluding remarks.

¹ <https://croderiv.ffzg.unizg.hr/>

2. CroDeriv

The development of CroDeriv took place in several phases. Its first version contained approximately 15,000 verbal lexemes collected from multiple online dictionaries. In this first phase of lexicon development, the primary goal was the analysis of the morphological structure of lexemes and the development of an appropriate data model that would enable queries over various morphological parameters across the lexicon. This approach proved valuable in many research areas, e.g., in the research of verbal aspect (Šojat et al., 2019), affixal meanings, affix ordering, combinations of particular affixes and roots as well as combinations of multiple affixes (Šojat et al., 2012; Šojat et al., 2013; Filko, 2020). However, the lexicon contained only lexemes of one part-of-speech (POS), and derivational links between them were not explicitly established. The morphological segmentation in this phase was performed semi-automatically using a naïve rule-based method, and the resulting segmentation was manually checked and corrected afterwards due to extensive allomorphy and phonological changes that take part at morpheme boundaries (e.g., assimilation or dropping) leading to a large error rate.

Morphological segmentation was based on the two-layered approach: the segmentation at the surface and the deep layer. Allomorphs were identified and marked for their type, i.e., root, prefix, suffix, or interfix, at the surface layer of analysis. Allomorphs were then associated with their representative morphs at the deep layer of presentation (e.g., *raščistiti* ‘to clear off’ < *raš-čist-i-ti* (surface layer) ~ *raz-čist-i-ti* (deep layer)). However, the segmentation in this first phase was shallow since lexemes were segmented to morphemes pertaining only to verbal derivational processes (verbal prefixes and suffixes), sometimes resulting in unsegmented complex stems. These unsegmented stems were morphological clusters inherited from non-verb derivational bases. To enable the grouping of lexemes into derivational families, i.e., ones that share the same lexical morpheme, “stems” were also associated with the root morpheme, albeit their content was often non-monomorphemic.

In the present stage, the data model has been completely revised in order to facilitate the expansion of the lexicon to other POS, mainly nouns and adjectives. This data model supports the full morphological decomposition of lexemes, regardless of their POS. At this stage, all data is analysed and segmented into morphemes manually, following the same two-layered approach. Besides, derivational links between lexemes are now explicitly established with identified stems (derivational bases), derivational affixes, and word-formation processes which take part in the derivation of a particular lexeme. Word-formation processes include prefixation, suffixation, simultaneous prefixation and suffixation, ablaut (usually combined with affixation), compounding (usually combined with affixation), etc. (see Filko et al., 2020, pp. 92–93; Šojat & Filko, 2023). Derivational affixes were also labelled for their morphosemantic meaning. Finally, lexemes were assigned their appropriate grammatical categories, i.e., aspect for verbs, gender for nouns, etc.

This design of the data model and the fact that CroDeriv now contains lexemes of all POS facilitates the representation of derivational families into graph-like structures. Derivational families are composed of lexemes that contain the same root, and in some cases, they can comprise more than 200 lexemes. The advantages of such representation authors have already described in Filko et al. (2021). These structures enable the

visual representation of derivational families and derivational paths from the lexical morpheme to a certain lexeme and thus make the Croatian derivational morphology more understandable and accessible to the general public.

3. Search and Visualization Interface

CroDeriv data is available to the users through a web application to be used in an internet browser. The architecture of the back-end system has already been described in Filko et al. (2020). Here we shall focus on the front-end interface, built as a Vue.js/Nuxt application. The interface is a part of a CroDeriv project website and it is designed to be intuitive and user-friendly, without the need for any specific usage instructions in the form of user documentation. In this way, we wanted to enable users who may not be technically savvy to also use the lexicon. However, all the features of the interface are described in place through thoughtfully positioned tooltips which contain enough information for users to be able to start using the lexicon right away.

Although we have constructed a special query language for searching the lexicon via API, the search interface eliminates the need for users to learn that language. Instead, the front-end application seamlessly translates users' input into the internal query language.

The search interface enables the user to search the CroDeriv lexicon in several ways: by lexemes, by morphological structures, and by word-formation patterns.

The simplest search is the one by the lexeme, in which the user can search for a particular lexeme or lexemes using the literal text input, or by using POSIX extended regular expressions (e.g., *rad.** will return all lexemes starting in *rad*, regardless of the morphological status of '*rad*' in their morphological structure: *rad* 'work', *raditi* 'to work', *rado* 'gladly', *radost* 'joy'...).

The search by morphological structure is facilitated by the visual search query builder. In this way, the user builds the morphological structure to be searched for against the lexicon database. The query builder allows the user to concatenate the search blocks representing morphemes into a sequence. With the possibility of using regular expressions along with the literal text inputs, the user can create very complex morphological queries, which is a feature that, to our best knowledge, makes this resource unique. For example, the user can search for all single or double prefixed lexemes, within a particular derivational family or across all families, or all lexemes with a particular suffix, or all compound lexemes, or any combination of these conditions. In Figure 1, the results of the search for all lexemes containing the prefix *pre-* followed by the root *√pis* are presented.

The word-formation search functions in a similar way as the morphological search, only with the difference that the building blocks of the search query represent the lexeme's derivational formants, i.e., stems and derivational affixes. This search is especially useful in searching for the same derivational patterns across derivational families or investigating the distribution of polysemous affixes.

All successful searches result in a list of lexemes in a tabular form that satisfy the search conditions (see Figure 1). The result lexeme list obtained by any of the ways of searching, can be additionally filtered by POS and relevant grammatical features. Additional properties associated with every result in this tabular view are lemma, part-of-speech, surface morphological structure, and word-formation pattern. For a more detailed description of every lexeme, the user can click on the link showing a detail page of the lexeme with a complete list of properties, which include deep morphological structure, word-formation process, and senses, i.e., morphosemantic meanings, of the affixes used in a derivation of a particular lexeme.

By clicking on the link for the graph, the user can get the visualization of the derivational path to the lexeme from its lexical root. The visualization is represented in the form of a labelled directed graph in which the nodes represent the lexemes in the derivational path and the edges represent the derivation links. Nodes are labelled with the lemma of the lexeme and the colour of the node denotes its POS. Derivational links are labelled with the name of the word-formation process as well as the sense of the derivational affix.

The interface shows a search for 'pre-' (prefix) and 'pis' (root). The results table is as follows:

LEMMA	POS	SURFACE LAYER	WORD FORMATION PATTERN	SEE DETAILS	SEE GRAPH
prepisati	V	pre - pis - a - ti	pre - pisati	SEE DETAILS	SEE GRAPH
prepisivač	N	pre - pis - iv - ač - ō	prepisivač - ač	SEE DETAILS	SEE GRAPH
prepisivačica	N	pre - pis - iv - ač - ic - a	prepisivač - ica	SEE DETAILS	SEE GRAPH
prepisivački	A	pre - pis - iv - ač - k - i	prepisivač - ki	SEE DETAILS	SEE GRAPH
prepisivan	T	pre - pis - iv - a - n	prepisivati - n	SEE DETAILS	SEE GRAPH
prepisivanje	N	pre - pis - iv - a - n - j - e	prepisivan - je	SEE DETAILS	SEE GRAPH

Fig. 1: Example of the search by morphological structure

4. Concluding Remarks

In this paper, we have presented the search and visualization interface for the Croatian morphological and derivational lexicon CroDeriv. This interface was built to make CroDeriv data available to a broad audience of users, from experts in the field, to students, Croatian language learners, and the general public. Most of the available morphological and derivational resources require a high level of technical expertise since they either have no user interface or utilize complex query languages. Our motivation was to leverage the advantages of a well-thought-out UX design to build an intuitive and easy-to-use, yet powerful interface. We believe that this interface can find its place both as a research platform, as well as a tool for teaching morphology in schools and universities or Croatian as a second and foreign language.

Acknowledgments

This paper and the work presented are partially supported by the Croatian National Consortium of CLARIN ERIC Research Infrastructure (HR-CLARIN).

References

- Filko, M. (2020). *Unutarleksičke i međuleksičke strukture imeničkoga dijela hrvatskoga leksika*. [Doctoral dissertation, Faculty of Humanities and Social Sciences, University of Zagreb]. ODRAZ.
- Filko, M., Šojat, K., & Štefanec, V. (2020). The Design of CroDeriv 2.0. *The Prague Bulletin of Mathematical Linguistics*, 115, 83–104. doi: 10.14712/00326585.006
- Filko, M., Šojat, K., & Štefanec, V. (2021). Deriving the Graph: Using Affixal Senses for Building Semantic Graphs. In F. Namer et al. (Eds.), *Proceedings of the Third International Workshop on Resources and Tools for Derivational Morphology (DeriMo 2021)* (pp. 120–128). ATILF (CNRS & UNIVERSITÉ DE LORRAINE).
- Kyánek, L. (2018). *Morphological resources of derivational word-formation relations*. [Technical report 61; ŮFAL, Charles University, Prague].
- Ljubešić, N., Klubička, F., Agić, Ž., & Jazbec, I.-P. (2016). New Inflectional Lexicons and Training Corpora for Improved Morphosyntactic Annotation of Croatian and Serbian. In N. Calzolari et al. (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)* (pp. 4264–4270). ELRA.
- Šojat, K., Srebačić, M., & Tadić, M. (2012). Derivational and Semantic Relations of Croatian Verbs. *Journal of Language Modelling*, 00(1), 111–142.
- Šojat, K., Srebačić, M., & Štefanec, V. (2013). CroDeriv i morfološka raščlamba hrvatskoga glagola. *Suvremena lingvistika*, 39(75), 75–96.
- Šojat, K., Kocijan, K., & Filko, M. (2019). Processing Croatian Aspectual Derivatives. In I. Mauro Mirto, M. Monteleone, & M. Silberztein (Eds.), *Formalizing Natural Languages with*

NooJ 2018 and Its Natural Language Processing Applications (pp. 50–61). Springer International Publishing.

Šojat, K., & Filko, M. (2023). Processing Croatian Morphology: Roots, Segmentation and Derivational Families. In K. Šojat, & M. Filko (Eds.), *Proceedings of the Fourth International Workshop on Resources and Tools for Derivational Morphology (DeriMo2023)* (pp. 61–70). Croatian Language Technologies Society.

Tadić, M., & Fulgosi, S. (2003). Building the Croatian Morphological Lexicon. In T. Erjavec, & D. Vitas (Eds.), *Proceedings of the EACL2003 Workshop on Morphological Processing of Slavic Languages* (pp. 41–46). Association for Computational Linguistics.

Contact information

Vanja Štefanec

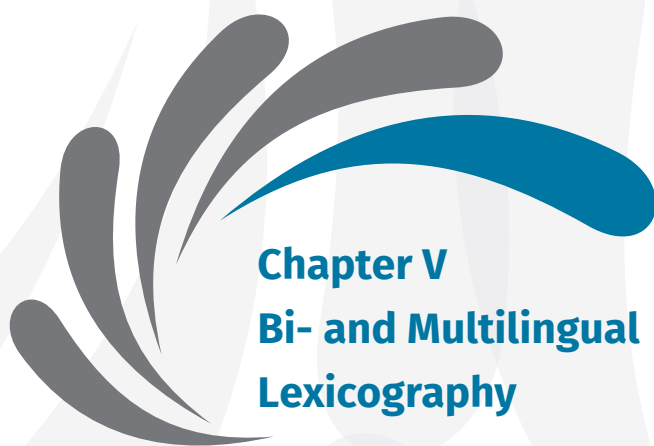
Faculty of Humanities and Social Sciences, University of Zagreb
vstefane@ffzg.unizg.hr

Krešimir Šojat

Faculty of Humanities and Social Sciences, University of Zagreb
ksojat@ffzg.unizg.hr

Matea Filko

Faculty of Humanities and Social Sciences, University of Zagreb
matea.filko@ffzg.unizg.hr



Chapter V
Bi- and Multilingual
Lexicography

Vladimír Benko, Zuzana Křihová, Boris Lehečka, and
Darina Vystrčilová

PERSIAN TO CZECH DICTIONARY

A Traditional Dictionary in the Era of AI?

Abstract We would like to introduce the results of the ELDI project (Electronic Lexical Database of Indo-Iranian Languages, Pilot module: Persian), launched in August 2020. One of the aims of the project was promoting the use of technologies in teaching languages. A website and a mobile application with the Persian–Czech dictionary were developed as the main planned results of the project. A new web-crawled Persian corpus was also created through cooperation with Comenius University in Bratislava and it was primarily used for the study and validation of lexical data, but it is now also directly linked to both applications and is open for use by students, teachers, and researchers in language studies and teaching. The participation of teachers from Charles University in Prague in the project will help to transfer the project outputs into practical teaching. This article also presents the results of a comparison of a dictionary and AI.

Keywords Persian; Czech; lexical database; web-crawled corpus; LLMs

1. Introduction

Persian is a language from the Indo-European language family that is spoken by almost 130 million people, mostly in Iran, Afghanistan, and Tajikistan, representing the three main language varieties of Persian, referred to as Farsi, Dari, and Tajik, respectively. While Farsi and Dari use a modified Arabic script, Tajik is written in modified Cyrillic. Although Czech is an Indo-European language as well, combining it with Persian in the same text presents an additional (typographic) challenge, as these languages differ in the direction of writing.

Our paper introduces some lexicographical and technical aspects of the ELDI Project, the aim of which was to launch a Persian–Czech dictionary website and mobile application (Vystrčilová et al., 2024a,b) and to promote new methods and technologies in teaching non-European languages. In addition, we present one (small-scale) experiment aimed at comparing experiences with traditional and AI-driven Persian to Czech translation.

2. The Lexical Data and the Methods Used to Process Them

The main data source for the Persian module of ELDI was a Persian lexical database that was originally compiled between 1996 and 2020 and served as the basis for three dictionaries that have already been published, two of them in paper form and one on-line (Vystrčilová, 2002; 2014; 2017). The lexical data originally collected on index cards

since 1996 were transferred into a text file and then into an Excel spreadsheet, where they were gradually organised into the structure that is described in Tables 1 and 2. After the start of the ELDI project in 2020, the data, divided into several packages, were transferred to *FLEX: FieldWorks Language Explorer* (SIL International, 2024.), where they were further processed, verified, and extended by a larger research team and shared using other tools developed by SIL International (2018): *LanguageDepot* and *FLEXBridge*.

At the end of the ELDI project in December 2023, a total of 57,317 dictionary entries were available in the FLEX database file shared by the project team, while 38,312 entries were made accessible in the website and the mobile application. Most of the entries were taken from previous work (52,199 processed entries plus a wordlist of the 24,661 most frequented lemmas) and during the project were subjected to further editing and verification.

The target number of entries is 50,000. The state of verification was the only criterion used for publication, while the frequency of occurrence of the Persian lemma in the mixed language corpora was used just to rank the search results in the ELDI website.

3. Sources of Lexical Evidence

Language corpora have been widely used as the main method for collecting and verifying lexical data since the late 1990s: the first ‘home-made’ Persian corpus, compiled from the texts of the *Hamshahri* newspaper – probably one of the first Persian open access media published in readable text form and giving readers access to the Persian script – was later replaced by *UPEC* (Uppsala Persian Corpus, Seraji et al., 2012) and the Persian section of *Open Subtitles* (Tiedemann, 2016), using the *WordSmith Tools* corpus manager (Lee et al., 2018; Scott, 2024). In preparation for the ELDI project in 2020, we began to use the *TalkBank Persian* corpus¹ (Rasooli et al., 2013), which could be accessed from the *Sketch Engine* website (Kilgariff et al., 2014) until April 2022, when its free use was discontinued following the conclusion of the *ELEXIS* Project. The need for a new Persian corpus, therefore, became apparent, and the most straightforward strategy appeared to be to create a new web-crawled corpus.

4. The Persian Web Corpus

During the last decade, the technology for creating corpora from web-crawled texts has been standardised. We were able to use FLOSS tools for most of the components in the processing pipeline (Baroni & Berdnardini, 2004; Pomikálek, 2011; Suchomel & Pomikálek, 2012; Michelfeit et al., 2014), and there were even tools available for the morpho-syntactic annotation of Persian texts, such as *TreeTagger* (Schmid, 1994), *UDPipe* (McDonald et al., 2013; Straka et al., 2016), and *CSTlemma* (Jongejan & Dalianis, 2009), all of which were trained on the *Seraji UD Treebank* (Seraji et al., 2016; Rasooli et al., 2022). As each of the aforementioned tools has its own specific drawbacks, we decided to combine their outputs in the hope of improving the overall

¹ See <https://www.sketchengine.eu/talkbank-persian-corpus/>; retrieved 21 July 2024.

quality of the annotations. The Persian web space turned out to be quite rich and after just a couple of days of crawling in April 2022 we had managed to create a corpus with over three billion tokens (Benko, 2022), which were processed using the standard pipeline created in the *Aranea* corpora Project.²

This corpus is now used by lexicographers as a primary source of lexical evidence for writing dictionary entries, and its 100-million sample is also linked to individual database entries and offers the user the possibility to view selected entries in the context of sentences in the *GDEX* setting (Kosem et al., 2019) or to study them in more detail using the *NoSketch Engine*³ Corpus Manager (Rychlý, 2007; Kilgarriff et al., 2014).

In the ELDI web application, the relative frequencies in the corpus are represented graphically with green squares, and the entries' frequency values are also used as one of the criteria for ordering the results of the dictionary search.

5. The Structure of the Lexicographic Data

The data structure is summarised in Figure 1 and Tables 1 and 2. Most of the fields come from the source database, but through the collaboration of the project team the database is undergoing further processing. In addition to adding the frequency of occurrence of the entries (at the lemma level) in the large corpora, especially in the newly created *Araneum Persicum*, more example sentences are included from the same source, and each sense definition is assigned to a higher semantic domain according to Moe (2012). More attention is paid to sorting the Czech equivalents into sense categories and their grammatical characteristics, and the system of notes is unified and accompanied by explanations and, for the most part, converted into metadata. Complex expressions (collocations, complex predicates, composites, idioms, etc.) are analysed and linked to relevant components. Chains considered to be variants of the main lexeme, especially stems of irregular verbs, Arabic plurals, complex predicates with alternative light verbs, or orthographic or colloquial variants, are linked to these lexemes. The participation of a native Persian speaker in the project editorial team has contributed significantly to extending the database with idioms and colloquialisms and to the verification of the Czech sense equivalents. Each sense definition is marked with a verification status, which is used to select data for the publicly accessible applications or as a guide for further processing.

Table 1: Structure of the dictionary entry. A) The headword

No	Field	Comments
1.	Main Headword	The basic form of the entry written in Persian script. This is the second priority field when sorting according to the Persian alphabet.
2.	Variants	Items shown as variants include the Present Stem of Irregular Verbs, Arabic Plurals, Complex Predicates with Alternative Light Verbs, Spelling and Colloquial Variants. The Main Headword and the Variants are represented as interconnected independent entries.

² See <http://aranea.juls.savba.sk/guest>, <http://unesco.uniba.sk/guest>; retrieved 21 July 2024.

³ See <https://nlp.fi.muni.cz/trac/noske>; retrieved 21 July 2024.

3.	Variant Type	Converted to FLEEx metadata.
4.	Pronunciation	Transcription of the Headword using the Czech alphabet. It is possible to use the pronunciation to search in the web app.
5.	Frequency	The letters M and T represent the relative frequencies in mixed corpora that meet current standards in terms of size, while the letters L and V represent data from two smaller corpora: UPEC (containing mostly literary language), and the Open Subtitles (spoken language). Six squares represent lexemes with a relative frequency > 1,000 ipm, 5 squares > 100 ipm, etc., and up to one square > 0.01 ipm.
6.	Components	Complex headwords can be divided into components, and seeing the connection between the components and complex forms helps the user to understand the meaning in the same way that the example sentences do.
7.	Complex Forms	Complex Predicates, Collocates, Compounds, Arabic Compounds, Derivatives, Idioms, Sayings.
8.	Links to Concordance	Link to GDEX sentences in a corpus.

The screenshot shows the ELDI web application interface. At the top, there is a search bar with the text 'Enter Query' and a search icon. Below the search bar, there is a navigation bar with 'Browse' and 'Search' links. The main content area displays the search results for the Persian word 'تعارف' (ta'arof). The entry is numbered 1. and includes a frequency indicator (6 squares). The sense definition is 'zdvořilé formální chování; zdvořilostní rituál, TA'ĀROF; etiketa; poklony; komplimenty; pobízení'. The Czech equivalents are listed as '4.3.7 Polite' and '4.3.4 Do good to'. The entry also includes a subentry for 'تعارفات دیپلماتیک' (diplomatic protocol) with the Czech equivalent 'diplomatický protokol'. The entry is also linked to 'analytický predikát: تعارف کردن' (analytical predicate: to offer, to invite (to enter)).

Fig. 1: Structure of the entry in the ELDI web application⁴

Table 2: Structure of the dictionary entry. B) Sense definition and Czech equivalents

No	Field	Comments
11.	Sense Number	
12.	Grammatical Information	Because of the large number of complex forms in Persian we consider the entry's semantic function rather than the part of speech. This field is crucial for identifying new senses when importing data into the FLEEx.
13.	Czech Equivalents (Definition in FLEEx)	The identified Czech Equivalents are organised in the Sense Definition Field in FLEEx. They may also contain context notes or indicate prepositional ties. Most of the other types of notes (e.g., register notes) were replaced by metadata taken from the FLEEx lists.
14.	Example Sentences	Example sentences compiled mainly from language corpora.

⁴ Reverse translation of the Czech senses: Sense 1: polite formal behaviour; courtesy ritual, TA'ĀROF; etiquette; paying respect; compliments; polite encouraging /inviting, Sense 2: getting to know each other. Complex expressions: تعارفات دیپلماتیک diplomatic protocol, تعارف کردن Sense 1: to offer, to invite (to enter).

15.	Example Translations	Czech translations of the example sentences.
16.	Prepositional Ties (Gram. Note in FLEx)	Prepositional ties and other grammatical and contextual notes. They appear at the beginning of the Definition line.
17.	Semantic Domains	Semantic Domain Thesaurus (Moe, 2012) created by Ron Moe incorporated into the FLEx is used.
18.	Register Notes (Usage in FLEx)	Note on usage in a specific register chosen from the FLEx list Usage.
19.	Hyperlinks to External Sources	Hyperlinks to external sources such as encyclopaedias, e.g., Wikipedia: The free encyclopedia (2024) or Encyclopædia Iranica (Yarshater et al. 1996–).
20.	Cross-references	References to related words or concepts within the dictionary, presenting links between synonyms, etc. (See Fig. 3)

There are other hidden fields in the lexical database that are used, for example, to facilitate the work of the editorial team and to select data for public presentation (Status), to set the alphabetical order or to prioritise search results during the search process (Lexeme Form, Reversals, and Allomorphs), to record sources of study and to verify the particular entry (Bibliography), and to record the history of data processing. The English equivalents were originally used only as a supplementary method of verifying the senses (comparing the English translation from Persian and from Czech), but it is now possible to search data using the sense definitions in English and using the English semantic domains in an advanced search.

6. Data Conversion

The validated data are exported from FLEx to the LIFT standard and then transformed into TEI Lex-0 format (Tasovac et al., 2024), which is being developed as a specialised version of the TEI standard for dictionaries. TEI Lex-0 is being developed by the DARIAH Working Group on Lexical Resources (2024).⁵ One goal of this standard is to capture different lexicographic practices in a uniform way and to allow for their comparison and for the mutual linking of existing dictionaries.

For sharing with other applications, the FLEx program exports data in its own XML format (Zook, 2015), with one of the options being the LIFT⁶ (Lexicon Interchange Format; SIL International, 2024b). The conversion from LIFT format to the TEI Lex-0 standard is fully automated, programmed in the XProc 3.0 programming language (XProc 3.0: Specifications, 2024). The MorganaXProc-IIIse application (Berndzen, 2024) is used to run it. The program converts elements of the LIFT format to corresponding elements of the TEI Lex-0 standard, assigns entries and senses with unique identifiers, and converts semantic category definitions (Moe, 2012) stored in a separate file to the <taxonomy> element in the header of the TEI Lex-0 document, etc. It also generates auxiliary data that the web application uses to present the dictionary, such as translations of data from the meta-language or SVG files, to represent the frequency of lemmas in the corpora.

⁵ See <https://github.com/DARIAH-ERIC/lexicalresources>; retrieved 21 July 2024.

⁶ See <https://github.com/sillsdev/lift-standard>; retrieved 21 July 2024.

The web application uses the XML database eXist-db (eXist Solutions, 2024) version 6.2.0. It comprises two separate components: one is designed for storing and indexing lexicographic data, the other provides access to the stored data through a user interface (web application) and a programming interface (REST API). Both components were created using the TEI Publisher 8.0 framework (e-editiones, 2024)⁷ and were later upgraded to version 9.0.⁸ This framework makes it possible to create the basic skeleton of the web application with common features, such as a full-text search, the option to browse a document by sections, or the ability to customise the appearance of the presented text using an ODD document.⁹ This basic skeleton of the web app can be extended using custom functions (in the XQuery programming language) or components (web pages, JavaScript functions, Web Components).¹⁰ TEI Publisher and related tools use open source code that can be modified.

Among the tools and features of the web application, the following are particularly useful: sorting search results according to the frequency of occurrence of a given lemma, facets enabling the further filtering of search results, superior semantic domains, the hypertext linking of all entries with the new Persian corpus, the mutual linking of orthographic or grammatical variants, components with complex expressions, and, for selected entries, hyperlinks to external encyclopaedic resources.

We paid great attention to the storage of lexicographic data and their indexing. The database contains many indexes in order to make the data and metadata efficiently searchable. The Persian–Czech dictionary in TEI Lex-0 format occupies 120 MB in one file. Indexing such a large file proved to take a disproportionately long time in the order of dozens of hours. This time was reduced by splitting the entire dictionary into separate files, with one file containing one entry. The data were uploaded to the server as a zip file, which also reduced the amount of time required to update the data.

The indexes and their evaluation during query processing affect what data are displayed in the first position as search results. Therefore, for example, when searching for a definition, the resulting score, which will affect the ranking of the results, is constructed based on several criteria scores: a) the frequency of the lemma in the new Persian corpus; b) the position of the searched equivalent in the meaning (equivalents at the top have higher relevance); c) the order of the sense in the entry, and d) the uniqueness of the sense (in which band among all the senses the search term is located).

The source code of the applications referred to above (the conversion of lexicographic data into the TEI Lex-0 standard and two eXist-db modules for data storage and indexing, as well as the user and programming interface) are freely available in the GitHub repository (Lehečka, 2024a-c).

⁷ See <https://github.com/eeditiones/tei-publisher-app/>; retrieved 21 July 2024.

⁸ Based on our experience working with the Electronic Database of Indo-Iranian Languages, we created a specialised version that offers features specific to dictionary data: TEI Lex-0 Publisher (Lehečka, 2024d).

⁹ ‘One Document Does It All’ (Text Encoding Initiative, 2024).

¹⁰ Web Components for TEI Publisher (2024), Mozilla (2024).

7. Enhancing Persian and Czech Translation: An Experiment Comparing the ELDI Persian–Czech Dictionary and ChatGPT 4

The integration of the ELDI Persian–Czech Dictionary (FaCsDict) into educational frameworks and translation practices marks a significant advancement in language pedagogy and translation precision. This part of the paper delves into the profound impact of this integration, offering a small-scale analysis of translations using FaCsDict versus those generated by OpenAI’s ChatGPT 4. The objective is to illuminate the capabilities and limitations of contemporary AI technologies in translating Persian and Czech and to demonstrate how these tools, when combined with human expertise, can refine translation methodologies and enhance educational outcomes.

The ELDI project is designed to address the underrepresentation of languages such as Persian and Czech in major lexical databases (LDBs). Many existing LDBs prioritise dominant languages like English, leading to a diminished capacity to represent culturally specific words and expressions in less widely spoken languages. By converting traditional dictionary data into a comprehensive Persian-based lexical database using FLEEx software, ELDI ensures a detailed and accurate representation of linguistic diversity. This transformation facilitates the preservation and effective translation of culturally specific terms, which is consistent with the goals of improving linguistic representation, as discussed, for example, by Giunchiglia et al. (2023).

7.1 The Efficiency and Contextual Appropriateness of ChatGPT 4

AI language models, including ChatGPT 4 (OpenAI, 2024), are highly proficient at generating translations rapidly, making them ideal for large-scale translation projects and real-time translation requirements. These models leverage advanced algorithms to consider the text’s context, producing translations that are not only fast but also more accurate and contextually appropriate. Recent advances in AI translation models have significantly enhanced multilingual text generation and understanding.

However, despite these advances, AI models often lack the nuanced understanding and cultural sensitivity that human translators possess. This deficiency can lead to potential inaccuracies or misinterpretations, especially in the case of specialised or technical terminology if their training data are not exhaustive (Grassini, 2023). Additionally, AI models often translate through an intermediate language, typically English, before converting the text into the target language. This process can introduce compounded errors, affecting both the intermediate English translation and the final output. For example, in practical translation tasks from Persian to Czech, it has been observed that AI systems may produce misleading or incorrect expressions in the Czech output due to inaccuracies introduced during the English intermediary stage. Such limitations underscore the necessity of integrating human expertise to refine and verify translations, ensuring cultural and contextual accuracy. Below is a sample that illustrates the compounded errors that can occur when AI models translate from Persian to Czech through an intermediate English stage, highlighting the need for human verification.

The specific instance involves the translation of the Persian compound verb **دل ریختن** [del rikhtan], meaning ‘to get scared’, which was inaccurately rendered. The English translation is acceptable, though not entirely accurate, but the Czech translation is misleading, as it is almost a literal rendition of the English version without considering Czech sentence structure and semantic nuances. The error lies not only in the translation of the idiom (lit. heart drop) into Czech but also in the rendition of the sentence structure into Czech. In Czech, the sentence actually means that the person is afraid that his father might come for him in the afternoon, rather than being afraid that his father won’t come and he will be left alone. The correct translation in Czech is shown in footnotes.

Sample Translation Issue: The following screenshot from ChatGPT 4 demonstrates the compounded errors:

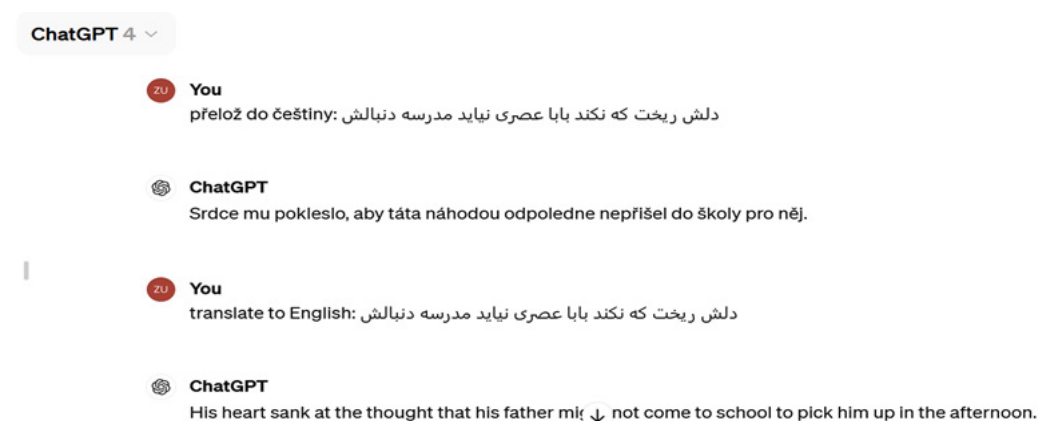


Fig. 2: Screenshot from ChatGPT 4 demonstrating the compounded errors.

To address these issues, the FaCsDict provides a correct translation directly from Persian to Czech, ensuring accurate and contextually appropriate translations. The FaCsDict approach bypasses the intermediate English translation, reducing the potential for compounded errors:

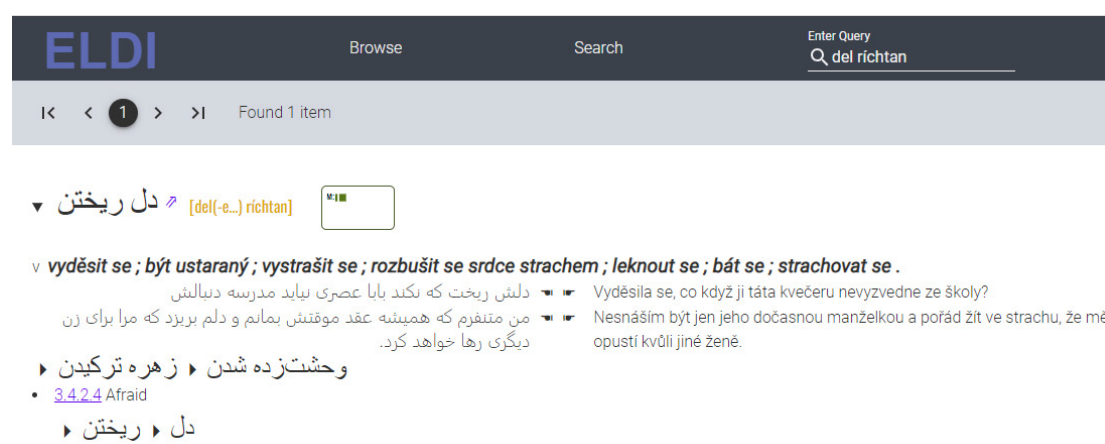


Fig. 3: Translation of the Persian idiom **دل ریختن** (lit. ‘heart sink’ by FaCsDict)¹¹

¹¹ Reverse translation of the Czech senses: to get scared; to be worried; to make the heart flutter with fear; to freak out; to fear; to worry. Example: She was scared, what if her dad doesn’t pick her up from school in the evening? I hate being just his temporary wife and always living in fear that he will leave me for another woman.

7.2 Challenges in Translating Polysemy and Idioms

One of the critical challenges in translation is handling polysemy and idioms effectively. AI models like ChatGPT 4, while proficient in generating coherent text, often struggle with these subtleties, leading to translations that may lack the depth and cultural accuracy required for literary texts. This challenge is evident in translations involving expressions with multiple meanings or idiomatic nuances.

For instance, ambiguities or inaccuracies often occur with grammatical constructs related to these issues, as exemplified by the Persian expression آب خوردن [āb khordan], which literally means ‘to drink water’. In Persian it also has a range of figurative meanings, which ChatGPT 4 captured in limited and incorrect ways. While the literal translations were accurate, all the figurative examples in Persian were incorrect, and their English translations were likewise inaccurate. In this case, the error made by ChatGPT 4 was more profound and problematic. ChatGPT 4 provided several example sentences by using an existing Persian idiom خوردن مثل آب [mesl-e āb khordan] and (easily) omitting the connection مثل [mesl] (like/as), without which the phrase is meaningless in Persian. Consequently, all the example sentences in Persian are incorrect, even though they seem to provide a comprehensive linguistic sample of the idiom studied. Additionally, while the literal translations were accurate, the figurative meanings that ChatGPT 4 presented in Persian were limited and incorrect. As a result, both the figurative examples in Persian and their English translations were inaccurate.



You

You are tasked with providing several semantically different usages of the Persian verb 'آب خوردن' (to drink water). You can conjugate the verb in different persons and tenses, but you must not change the verb part of the compound verb. Use both the literal and figurative meaning of the verb.



ChatGPT

Sure! Here are several semantically different usages of the Persian verb 'آب خوردن' (to drink water), utilizing both literal and figurative meanings, and conjugated in different persons and tenses:

Literal Meanings

1. Present Tense (Third Person Singular)

- **Sentence:** او هر روز بعد از ورزش آب می‌خورد.
- **Translation:** He drinks water every day after exercise.

2. Past Tense (First Person Singular)

- **Sentence:** من دیروز بعد از پیاده‌روی طولانی آب خوردم.

Fig. 4: Screenshot of the ChatGPT 4 translation of the phrase آب خوردن

FaCsDict addresses these complexities by providing contextually rich examples and detailed semantic classifications, ensuring that each word's nuanced meanings are accurately captured and translated. This capability is demonstrated in the following slide, which showcases FaCsDict's effectiveness in accurately translating figurative expressions, highlighting its handling of polysemy and idiomatic phrases compared to AI translation.

ELDI Browse Search Enter Query Q áb chordan

Found 2 items

▼ آب خوردن ¹ ² ³ ⁴ ⁵ ⁶ ⁷ ⁸ ⁹ ¹⁰ ¹¹ ¹² ¹³ ¹⁴ ¹⁵ ¹⁶ ¹⁷ ¹⁸ ¹⁹ ²⁰ ²¹ ²² ²³ ²⁴ ²⁵ ²⁶ ²⁷ ²⁸ ²⁹ ³⁰ ³¹ ³² ³³ ³⁴ ³⁵ ³⁶ ³⁷ ³⁸ ³⁹ ⁴⁰ ⁴¹ ⁴² ⁴³ ⁴⁴ ⁴⁵ ⁴⁶ ⁴⁷ ⁴⁸ ⁴⁹ ⁵⁰ ⁵¹ ⁵² ⁵³ ⁵⁴ ⁵⁵ ⁵⁶ ⁵⁷ ⁵⁸ ⁵⁹ ⁶⁰ ⁶¹ ⁶² ⁶³ ⁶⁴ ⁶⁵ ⁶⁶ ⁶⁷ ⁶⁸ ⁶⁹ ⁷⁰ ⁷¹ ⁷² ⁷³ ⁷⁴ ⁷⁵ ⁷⁶ ⁷⁷ ⁷⁸ ⁷⁹ ⁸⁰ ⁸¹ ⁸² ⁸³ ⁸⁴ ⁸⁵ ⁸⁶ ⁸⁷ ⁸⁸ ⁸⁹ ⁹⁰ ⁹¹ ⁹² ⁹³ ⁹⁴ ⁹⁵ ⁹⁶ ⁹⁷ ⁹⁸ ⁹⁹ ¹⁰⁰ ¹⁰¹ ¹⁰² ¹⁰³ ¹⁰⁴ ¹⁰⁵ ¹⁰⁶ ¹⁰⁷ ¹⁰⁸ ¹⁰⁹ ¹¹⁰ ¹¹¹ ¹¹² ¹¹³ ¹¹⁴ ¹¹⁵ ¹¹⁶ ¹¹⁷ ¹¹⁸ ¹¹⁹ ¹²⁰ ¹²¹ ¹²² ¹²³ ¹²⁴ ¹²⁵ ¹²⁶ ¹²⁷ ¹²⁸ ¹²⁹ ¹³⁰ ¹³¹ ¹³² ¹³³ ¹³⁴ ¹³⁵ ¹³⁶ ¹³⁷ ¹³⁸ ¹³⁹ ¹⁴⁰ ¹⁴¹ ¹⁴² ¹⁴³ ¹⁴⁴ ¹⁴⁵ ¹⁴⁶ ¹⁴⁷ ¹⁴⁸ ¹⁴⁹ ¹⁵⁰ ¹⁵¹ ¹⁵² ¹⁵³ ¹⁵⁴ ¹⁵⁵ ¹⁵⁶ ¹⁵⁷ ¹⁵⁸ ¹⁵⁹ ¹⁶⁰ ¹⁶¹ ¹⁶² ¹⁶³ ¹⁶⁴ ¹⁶⁵ ¹⁶⁶ ¹⁶⁷ ¹⁶⁸ ¹⁶⁹ ¹⁷⁰ ¹⁷¹ ¹⁷² ¹⁷³ ¹⁷⁴ ¹⁷⁵ ¹⁷⁶ ¹⁷⁷ ¹⁷⁸ ¹⁷⁹ ¹⁸⁰ ¹⁸¹ ¹⁸² ¹⁸³ ¹⁸⁴ ¹⁸⁵ ¹⁸⁶ ¹⁸⁷ ¹⁸⁸ ¹⁸⁹ ¹⁹⁰ ¹⁹¹ ¹⁹² ¹⁹³ ¹⁹⁴ ¹⁹⁵ ¹⁹⁶ ¹⁹⁷ ¹⁹⁸ ¹⁹⁹ ²⁰⁰ ²⁰¹ ²⁰² ²⁰³ ²⁰⁴ ²⁰⁵ ²⁰⁶ ²⁰⁷ ²⁰⁸ ²⁰⁹ ²¹⁰ ²¹¹ ²¹² ²¹³ ²¹⁴ ²¹⁵ ²¹⁶ ²¹⁷ ²¹⁸ ²¹⁹ ²²⁰ ²²¹ ²²² ²²³ ²²⁴ ²²⁵ ²²⁶ ²²⁷ ²²⁸ ²²⁹ ²³⁰ ²³¹ ²³² ²³³ ²³⁴ ²³⁵ ²³⁶ ²³⁷ ²³⁸ ²³⁹ ²⁴⁰ ²⁴¹ ²⁴² ²⁴³ ²⁴⁴ ²⁴⁵ ²⁴⁶ ²⁴⁷ ²⁴⁸ ²⁴⁹ ²⁵⁰ ²⁵¹ ²⁵² ²⁵³ ²⁵⁴ ²⁵⁵ ²⁵⁶ ²⁵⁷ ²⁵⁸ ²⁵⁹ ²⁶⁰ ²⁶¹ ²⁶² ²⁶³ ²⁶⁴ ²⁶⁵ ²⁶⁶ ²⁶⁷ ²⁶⁸ ²⁶⁹ ²⁷⁰ ²⁷¹ ²⁷² ²⁷³ ²⁷⁴ ²⁷⁵ ²⁷⁶ ²⁷⁷ ²⁷⁸ ²⁷⁹ ²⁸⁰ ²⁸¹ ²⁸² ²⁸³ ²⁸⁴ ²⁸⁵ ²⁸⁶ ²⁸⁷ ²⁸⁸ ²⁸⁹ ²⁹⁰ ²⁹¹ ²⁹² ²⁹³ ²⁹⁴ ²⁹⁵ ²⁹⁶ ²⁹⁷ ²⁹⁸ ²⁹⁹ ³⁰⁰ ³⁰¹ ³⁰² ³⁰³ ³⁰⁴ ³⁰⁵ ³⁰⁶ ³⁰⁷ ³⁰⁸ ³⁰⁹ ³¹⁰ ³¹¹ ³¹² ³¹³ ³¹⁴ ³¹⁵ ³¹⁶ ³¹⁷ ³¹⁸ ³¹⁹ ³²⁰ ³²¹ ³²² ³²³ ³²⁴ ³²⁵ ³²⁶ ³²⁷ ³²⁸ ³²⁹ ³³⁰ ³³¹ ³³² ³³³ ³³⁴ ³³⁵ ³³⁶ ³³⁷ ³³⁸ ³³⁹ ³⁴⁰ ³⁴¹ ³⁴² ³⁴³ ³⁴⁴ ³⁴⁵ ³⁴⁶ ³⁴⁷ ³⁴⁸ ³⁴⁹ ³⁵⁰ ³⁵¹ ³⁵² ³⁵³ ³⁵⁴ ³⁵⁵ ³⁵⁶ ³⁵⁷ ³⁵⁸ ³⁵⁹ ³⁶⁰ ³⁶¹ ³⁶² ³⁶³ ³⁶⁴ ³⁶⁵ ³⁶⁶ ³⁶⁷ ³⁶⁸ ³⁶⁹ ³⁷⁰ ³⁷¹ ³⁷² ³⁷³ ³⁷⁴ ³⁷⁵ ³⁷⁶ ³⁷⁷ ³⁷⁸ ³⁷⁹ ³⁸⁰ ³⁸¹ ³⁸² ³⁸³ ³⁸⁴ ³⁸⁵ ³⁸⁶ ³⁸⁷ ³⁸⁸ ³⁸⁹ ³⁹⁰ ³⁹¹ ³⁹² ³⁹³ ³⁹⁴ ³⁹⁵ ³⁹⁶ ³⁹⁷ ³⁹⁸ ³⁹⁹ ⁴⁰⁰ ⁴⁰¹ ⁴⁰² ⁴⁰³ ⁴⁰⁴ ⁴⁰⁵ ⁴⁰⁶ ⁴⁰⁷ ⁴⁰⁸ ⁴⁰⁹ ⁴¹⁰ ⁴¹¹ ⁴¹² ⁴¹³ ⁴¹⁴ ⁴¹⁵ ⁴¹⁶ ⁴¹⁷ ⁴¹⁸ ⁴¹⁹ ⁴²⁰ ⁴²¹ ⁴²² ⁴²³ ⁴²⁴ ⁴²⁵ ⁴²⁶ ⁴²⁷ ⁴²⁸ ⁴²⁹ ⁴³⁰ ⁴³¹ ⁴³² ⁴³³ ⁴³⁴ ⁴³⁵ ⁴³⁶ ⁴³⁷ ⁴³⁸ ⁴³⁹ ⁴⁴⁰ ⁴⁴¹ ⁴⁴² ⁴⁴³ ⁴⁴⁴ ⁴⁴⁵ ⁴⁴⁶ ⁴⁴⁷ ⁴⁴⁸ ⁴⁴⁹ ⁴⁵⁰ ⁴⁵¹ ⁴⁵² ⁴⁵³ ⁴⁵⁴ ⁴⁵⁵ ⁴⁵⁶ ⁴⁵⁷ ⁴⁵⁸ ⁴⁵⁹ ⁴⁶⁰ ⁴⁶¹ ⁴⁶² ⁴⁶³ ⁴⁶⁴ ⁴⁶⁵ ⁴⁶⁶ ⁴⁶⁷ ⁴⁶⁸ ⁴⁶⁹ ⁴⁷⁰ ⁴⁷¹ ⁴⁷² ⁴⁷³ ⁴⁷⁴ ⁴⁷⁵ ⁴⁷⁶ ⁴⁷⁷ ⁴⁷⁸ ⁴⁷⁹ ⁴⁸⁰ ⁴⁸¹ ⁴⁸² ⁴⁸³ ⁴⁸⁴ ⁴⁸⁵ ⁴⁸⁶ ⁴⁸⁷ ⁴⁸⁸ ⁴⁸⁹ ⁴⁹⁰ ⁴⁹¹ ⁴⁹² ⁴⁹³ ⁴⁹⁴ ⁴⁹⁵ ⁴⁹⁶ ⁴⁹⁷ ⁴⁹⁸ ⁴⁹⁹ ⁵⁰⁰ ⁵⁰¹ ⁵⁰² ⁵⁰³ ⁵⁰⁴ ⁵⁰⁵ ⁵⁰⁶ ⁵⁰⁷ ⁵⁰⁸ ⁵⁰⁹ ⁵¹⁰ ⁵¹¹ ⁵¹² ⁵¹³ ⁵¹⁴ ⁵¹⁵ ⁵¹⁶ ⁵¹⁷ ⁵¹⁸ ⁵¹⁹ ⁵²⁰ ⁵²¹ ⁵²² ⁵²³ ⁵²⁴ ⁵²⁵ ⁵²⁶ ⁵²⁷ ⁵²⁸ ⁵²⁹ ⁵³⁰ ⁵³¹ ⁵³² ⁵³³ ⁵³⁴ ⁵³⁵ ⁵³⁶ ⁵³⁷ ⁵³⁸ ⁵³⁹ ⁵⁴⁰ ⁵⁴¹ ⁵⁴² ⁵⁴³ ⁵⁴⁴ ⁵⁴⁵ ⁵⁴⁶ ⁵⁴⁷ ⁵⁴⁸ ⁵⁴⁹ ⁵⁵⁰ ⁵⁵¹ ⁵⁵² ⁵⁵³ ⁵⁵⁴ ⁵⁵⁵ ⁵⁵⁶ ⁵⁵⁷ ⁵⁵⁸ ⁵⁵⁹ ⁵⁶⁰ ⁵⁶¹ ⁵⁶² ⁵⁶³ ⁵⁶⁴ ⁵⁶⁵ ⁵⁶⁶ ⁵⁶⁷ ⁵⁶⁸ ⁵⁶⁹ ⁵⁷⁰ ⁵⁷¹ ⁵⁷² ⁵⁷³ ⁵⁷⁴ ⁵⁷⁵ ⁵⁷⁶ ⁵⁷⁷ ⁵⁷⁸ ⁵⁷⁹ ⁵⁸⁰ ⁵⁸¹ ⁵⁸² ⁵⁸³ ⁵⁸⁴ ⁵⁸⁵ ⁵⁸⁶ ⁵⁸⁷ ⁵⁸⁸ ⁵⁸⁹ ⁵⁹⁰ ⁵⁹¹ ⁵⁹² ⁵⁹³ ⁵⁹⁴ ⁵⁹⁵ ⁵⁹⁶ ⁵⁹⁷ ⁵⁹⁸ ⁵⁹⁹ ⁶⁰⁰ ⁶⁰¹ ⁶⁰² ⁶⁰³ ⁶⁰⁴ ⁶⁰⁵ ⁶⁰⁶ ⁶⁰⁷ ⁶⁰⁸ ⁶⁰⁹ ⁶¹⁰ ⁶¹¹ ⁶¹² ⁶¹³ ⁶¹⁴ ⁶¹⁵ ⁶¹⁶ ⁶¹⁷ ⁶¹⁸ ⁶¹⁹ ⁶²⁰ ⁶²¹ ⁶²² ⁶²³ ⁶²⁴ ⁶²⁵ ⁶²⁶ ⁶²⁷ ⁶²⁸ ⁶²⁹ ⁶³⁰ ⁶³¹ ⁶³² ⁶³³ ⁶³⁴ ⁶³⁵ ⁶³⁶ ⁶³⁷ ⁶³⁸ ⁶³⁹ ⁶⁴⁰ ⁶⁴¹ ⁶⁴² ⁶⁴³ ⁶⁴⁴ ⁶⁴⁵ ⁶⁴⁶ ⁶⁴⁷ ⁶⁴⁸ ⁶⁴⁹ ⁶⁵⁰ ⁶⁵¹ ⁶⁵² ⁶⁵³ ⁶⁵⁴ ⁶⁵⁵ ⁶⁵⁶ ⁶⁵⁷ ⁶⁵⁸ ⁶⁵⁹ ⁶⁶⁰ ⁶⁶¹ ⁶⁶² ⁶⁶³ ⁶⁶⁴ ⁶⁶⁵ ⁶⁶⁶ ⁶⁶⁷ ⁶⁶⁸ ⁶⁶⁹ ⁶⁷⁰ ⁶⁷¹ ⁶⁷² ⁶⁷³ ⁶⁷⁴ ⁶⁷⁵ ⁶⁷⁶ ⁶⁷⁷ ⁶⁷⁸ ⁶⁷⁹ ⁶⁸⁰ ⁶⁸¹ ⁶⁸² ⁶⁸³ ⁶⁸⁴ ⁶⁸⁵ ⁶⁸⁶ ⁶⁸⁷ ⁶⁸⁸ ⁶⁸⁹ ⁶⁹⁰ ⁶⁹¹ ⁶⁹² ⁶⁹³ ⁶⁹⁴ ⁶⁹⁵ ⁶⁹⁶ ⁶⁹⁷ ⁶⁹⁸ ⁶⁹⁹ ⁷⁰⁰ ⁷⁰¹ ⁷⁰² ⁷⁰³ ⁷⁰⁴ ⁷⁰⁵ ⁷⁰⁶ ⁷⁰⁷ ⁷⁰⁸ ⁷⁰⁹ ⁷¹⁰ ⁷¹¹ ⁷¹² ⁷¹³ ⁷¹⁴ ⁷¹⁵ ⁷¹⁶ ⁷¹⁷ ⁷¹⁸ ⁷¹⁹ ⁷²⁰ ⁷²¹ ⁷²² ⁷²³ ⁷²⁴ ⁷²⁵ ⁷²⁶ ⁷²⁷ ⁷²⁸ ⁷²⁹ ⁷³⁰ ⁷³¹ ⁷³² ⁷³³ ⁷³⁴ ⁷³⁵ ⁷³⁶ ⁷³⁷ ⁷³⁸ ⁷³⁹ ⁷⁴⁰ ⁷⁴¹ ⁷⁴² ⁷⁴³ ⁷⁴⁴ ⁷⁴⁵ ⁷⁴⁶ ⁷⁴⁷ ⁷⁴⁸ ⁷⁴⁹ ⁷⁵⁰ ⁷⁵¹ ⁷⁵² ⁷⁵³ ⁷⁵⁴ ⁷⁵⁵ ⁷⁵⁶ ⁷⁵⁷ ⁷⁵⁸ ⁷⁵⁹ ⁷⁶⁰ ⁷⁶¹ ⁷⁶² ⁷⁶³ ⁷⁶⁴ ⁷⁶⁵ ⁷⁶⁶ ⁷⁶⁷ ⁷⁶⁸ ⁷⁶⁹ ⁷⁷⁰ ⁷⁷¹ ⁷⁷² ⁷⁷³ ⁷⁷⁴ ⁷⁷⁵ ⁷⁷⁶ ⁷⁷⁷ ⁷⁷⁸ ⁷⁷⁹ ⁷⁸⁰ ⁷⁸¹ ⁷⁸² ⁷⁸³ ⁷⁸⁴ ⁷⁸⁵ ⁷⁸⁶ ⁷⁸⁷ ⁷⁸⁸ ⁷⁸⁹ ⁷⁹⁰ ⁷⁹¹ ⁷⁹² ⁷⁹³ ⁷⁹⁴ ⁷⁹⁵ ⁷⁹⁶ ⁷⁹⁷ ⁷⁹⁸ ⁷⁹⁹ ⁸⁰⁰ ⁸⁰¹ ⁸⁰² ⁸⁰³ ⁸⁰⁴ ⁸⁰⁵ ⁸⁰⁶ ⁸⁰⁷ ⁸⁰⁸ ⁸⁰⁹ ⁸¹⁰ ⁸¹¹ ⁸¹² ⁸¹³ ⁸¹⁴ ⁸¹⁵ ⁸¹⁶ ⁸¹⁷ ⁸¹⁸ ⁸¹⁹ ⁸²⁰ ⁸²¹ ⁸²² ⁸²³ ⁸²⁴ ⁸²⁵ ⁸²⁶ ⁸²⁷ ⁸²⁸ ⁸²⁹ ⁸³⁰ ⁸³¹ ⁸³² ⁸³³ ⁸³⁴ ⁸³⁵ ⁸³⁶ ⁸³⁷ ⁸³⁸ ⁸³⁹ ⁸⁴⁰ ⁸⁴¹ ⁸⁴² ⁸⁴³ ⁸⁴⁴ ⁸⁴⁵ ⁸⁴⁶ ⁸⁴⁷ ⁸⁴⁸ ⁸⁴⁹ ⁸⁵⁰ ⁸⁵¹ ⁸⁵² ⁸⁵³ ⁸⁵⁴ ⁸⁵⁵ ⁸⁵⁶ ⁸⁵⁷ ⁸⁵⁸ ⁸⁵⁹ ⁸⁶⁰ ⁸⁶¹ ⁸⁶² ⁸⁶³ ⁸⁶⁴ ⁸⁶⁵ ⁸⁶⁶ ⁸⁶⁷ ⁸⁶⁸ ⁸⁶⁹ ⁸⁷⁰ ⁸⁷¹ ⁸⁷² ⁸⁷³ ⁸⁷⁴ ⁸⁷⁵ ⁸⁷⁶ ⁸⁷⁷ ⁸⁷⁸ ⁸⁷⁹ ⁸⁸⁰ ⁸⁸¹ ⁸⁸² ⁸⁸³ ⁸⁸⁴ ⁸⁸⁵ ⁸⁸⁶ ⁸⁸⁷ ⁸⁸⁸ ⁸⁸⁹ ⁸⁹⁰ ⁸⁹¹ ⁸⁹² ⁸⁹³ ⁸⁹⁴ ⁸⁹⁵ ⁸⁹⁶ ⁸⁹⁷ ⁸⁹⁸ ⁸⁹⁹ ⁹⁰⁰ ⁹⁰¹ ⁹⁰² ⁹⁰³ ⁹⁰⁴ ⁹⁰⁵ ⁹⁰⁶ ⁹⁰⁷ ⁹⁰⁸ ⁹⁰⁹ ⁹¹⁰ ⁹¹¹ ⁹¹² ⁹¹³ ⁹¹⁴ ⁹¹⁵ ⁹¹⁶ ⁹¹⁷ ⁹¹⁸ ⁹¹⁹ ⁹²⁰ ⁹²¹ ⁹²² ⁹²³ ⁹²⁴ ⁹²⁵ ⁹²⁶ ⁹²⁷ ⁹²⁸ ⁹²⁹ ⁹³⁰ ⁹³¹ ⁹³² ⁹³³ ⁹³⁴ ⁹³⁵ ⁹³⁶ ⁹³⁷ ⁹³⁸ ⁹³⁹ ⁹⁴⁰ ⁹⁴¹ ⁹⁴² ⁹⁴³ ⁹⁴⁴ ⁹⁴⁵ ⁹⁴⁶ ⁹⁴⁷ ⁹⁴⁸ ⁹⁴⁹ ⁹⁵⁰ ⁹⁵¹ ⁹⁵² ⁹⁵³ ⁹⁵⁴ ⁹⁵⁵ ⁹⁵⁶ ⁹⁵⁷ ⁹⁵⁸ ⁹⁵⁹ ⁹⁶⁰ ⁹⁶¹ ⁹⁶² ⁹⁶³ ⁹⁶⁴ ⁹⁶⁵ ⁹⁶⁶ ⁹⁶⁷ ⁹⁶⁸ ⁹⁶⁹ ⁹⁷⁰ ⁹⁷¹ ⁹⁷² ⁹⁷³ ⁹⁷⁴ ⁹⁷⁵ ⁹⁷⁶ ⁹⁷⁷ ⁹⁷⁸ ⁹⁷⁹ ⁹⁸⁰ ⁹⁸¹ ⁹⁸² ⁹⁸³ ⁹⁸⁴ ⁹⁸⁵ ⁹⁸⁶ ⁹⁸⁷ ⁹⁸⁸ ⁹⁸⁹ ⁹⁹⁰ ⁹⁹¹ ⁹⁹² ⁹⁹³ ⁹⁹⁴ ⁹⁹⁵ ⁹⁹⁶ ⁹⁹⁷ ⁹⁹⁸ ⁹⁹⁹ ¹⁰⁰⁰

Fig. 5: Translation of the verb/idiom آب خوردن and related complex expressions by FaCsDict¹²

The integration of human expertise is crucial for overcoming the limitations of AI translations. While AI models are adept at handling large volumes of text and providing quick translations, human translators bring the essential skills of cultural sensitivity and contextual understanding. This combination is especially important for languages with less representation in AI training datasets, like Persian and Czech. For instance, idiomatic expressions and nuanced grammatical structures often require a deep understanding of both the source and target cultures, which AI models might not fully possess due to their reliance on statistical patterns rather than lived cultural experiences.

AI models, including ChatGPT 4, are utilised for initial, preliminary translations because of their speed and contextual capabilities. However, the refinement process is supported by FaCsDict, which provides contextually rich examples and detailed semantic classifications. This hybrid approach ensures both efficiency and accuracy, preserving the literary quality of the translated text by accurately capturing the nuanced meanings of polysemous terms and idiomatic expressions. While this hybrid approach ensures both efficiency and accuracy, preserving the literary quality of the translated text by accurately capturing the nuanced meanings of polysemous terms and idiomatic expressions, it should be noted that FaCsDict is still under development and continually being improved. Additionally, using corpus tools is highly effective for verifying the correctness of idiomatic meanings.

¹² Reverse translation of the Czech senses: 1. to drink (water). 2. to originate (from); to come from. Example: I don't know where the rumours are coming from. 3. to cost. The trip cost me a fortune. Complexes: a) watering hole; trough for watering cattle, b) simple; easy; like nothing; a piece of cake.

8. Conclusion

During the ELDI project, a Persian–Czech Database containing some 52,000 processed entries compiled over the period 1996–2020 was imported into a shared FieldWorks Language Explorer file and further extended, modified, and verified by the project team. Over 38,000 dictionary entries chosen by the status of verification were converted into two publicly accessible applications in June 2024: a website using TEI Lex-0 based on TEI Publisher and a mobile app based on Dictionary App Builder. The following features are considered to be its top advantages: content rich in idioms; the direct linking of dictionary entries to GDEX sentences in the corpus Araneum Persicum; solutions to some of the intricacies of Persian orthography; and the inclusion of frequency data to improve the search methods and the ordering of results; the semantic domains assigned to each sense definition, which in combination with the facets (Figure 3 right) and search tools provide users with new ways to explore vocabulary and organize the lexical data.

Significant side results of the ELDI project are a new Persian language corpus called Araneum Persicum and software based on TEI Publisher that is ready for use with other languages. The use of the project results in the teaching of languages is being promoted by the team members at Charles University. Based on a small-scale experiment, the most effective method for educational purposes and translation practices (demonstrated through the example of a literary text in Persian and Czech) involves leveraging the strengths of both AI models and the FaCsDict. The key members of the team are continuing to collaborate on further editorial work as well as consulting on further modifications of the language corpus and the web application.

References

- Baroni, M., & Bernardini, S. (2004). BootCaT: Bootstrapping Corpora and Terms from the Web. In *Proceedings of LREC 2004* (pp. 1313–1316).
- Benko, V. (2022). Aranea Go Middle East: Persicum. In A. Horák, P. Rychlý, & A. Rambousek (Eds.), *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2022* (pp. 1–9). Tribun.
- Berndzen, A. (2024). *MorganaXProc-IIIse*. Retrieved July 21, 2024, from <https://www.xml-project.com/morganaxproc-iiiise.html>
- DARIAH Working Group on Lexical Resources. Retrieved July, 21 2024, from <https://www.dariah.eu/activities/working-groups/lexical-resources/>
- eXist Solutions. (2024). *eXist-db*. Retrieved July, 21, 2024, from <https://exist-db.org>
- FLEX: FieldWorks Language Explorer. Retrieved July, 21, 2024, from <https://software.sil.org/fieldworks/>
- Giunchiglia, F., Bella, G., Chandran Nair, N., Chi, Y., & Xu, H. (2023). Representing Interlingual Meaning in Lexical Databases. *Artificial Intelligence Review*, 56, 1–17. <https://arxiv.org/abs/2301.09169>

Grassini, S. (2023). Shaping the Future of Education: Exploring the Potential and Consequences of AI and ChatGPT in Educational Settings. *Education Sciences* 13(7), 692. <https://doi.org/10.3390/educsci13070692>

Jongejan, B., & Dalianis, H. (2009). Automatic training of lemmatization rules that handle morphological changes in pre-, in- and suffixes alike. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP* (pp. 145–153). Association for Computational Linguistics.

Kosem, I., Koppel, K., Zingano Kuhn, T., Michelfeit, J., & Tiberius, C. (2019). Identification and automatic extraction of good dictionary examples: the case(s) of GDEX. *International Journal of Lexicography*, 32(2), 119–137.

Lee, S. M., Scott, M., & Shin, D. (2018). *A Practical Approach to Corpus-based Analysis: Introducing WordSmith Tools, Range, and Other Packages*. Lehečka, B. (2024a). *LeDIIR Users*. Retrieved July, 21, 2024, from <https://github.com/daliboris/lediir-web-users/>

Lehečka, B. (2024b). *LeDIIR Web Application*. Retrieved July, 21, 2024, from <https://github.com/daliboris/lediir-web-app/>

Lehečka, B. (2024c). *LeDIIR Web Data*. Retrieved July, 21, 2024, from <https://github.com/daliboris/lediir-web-data/>

Lehečka, B. (2024d). *TEI Lex-0 Publisher*. Version 9.0.1. Retrieved July, 21, 2024, from <https://github.com/DARIAH-ERIC/teilex0-publisher>

McDonald, R., Nivre, J., Quirnbach-Brundage, Y., Goldberg, Z., Das, D., Ganchev, K., Hall, K., Petrov, S., Zhang, H., Täckström, O., Bedini, C., Bertomeu Castelló, N., & Lee, J. (2013). Universal Dependency Annotation for Multilingual Parsing. In *Proceedings of ACL* (pp. 92–97). Association for Computational Linguistics.

Michelfeit, J., Pomikálek, J., & Suchomel, V. (2014). Text Tokenisation Using unitok. In *8th Workshop on Recent Advances in Slavonic Natural Language Processing* (pp. 71–75). Tribun EU.

Moe, R. (2012). *Semantic Domains*. Retrieved July, 21, 2024, from <https://semdom.org>

Mozilla. (2024). *Web Components*. Retrieved July, 21, 2024, from https://developer.mozilla.org/en-US/docs/Web/API/Web_components

OpenAI. (2024). *ChatGPT 4*. Retrieved May, 28, 2024, from <https://chat.openai.com>

Pomikálek, J. (2011). *Removing boilerplate and duplicate content from web corpora*. [Doctoral thesis, Masaryk university, Faculty of informatics].

Rasooli, M. S., Kouhestani, M., & Moloodi, A. (2013). Development of a Persian syntactic dependency treebank. In *Proceedings of NAACL-HLT* (pp. 306–314). Association for Computational Linguistics.

Rychlý, P. (2007). Manatee/Bonito – A Modular Corpus Manager. In *RASLAN* (pp. 65–70). Masaryk University.

- Safari, P., Rasooli, M. S., Moloodi, A., & Nourian, A. (2022). The Persian Dependency Treebank Made Universal. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 7078–7087). European Language Resources Association.
- Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing* (pp. 44–49).
- Scott, M. (2008). *WordSmith Tools version 5*. Lexical Analysis Software Ltd. Retrieved July 21, 2024, from <http://www.lexically.net/wordsmith/index.html>
- Scott, M. (2024). *WordSmith Tools version 9* (64 bit version). Lexical Analysis Software.
- Seraji, M., Ginter, F., & Nivre, J. (2016). Universal Dependencies for Persian. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LRE)* (pp. 2361–2365). European Language Resources Association.
- Seraji, M., Megyesi, B., & Nivre, J. (2012). A Basic Language Resource Kit for Persian. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)* (pp. 2245–2252). European Language Resources Association.
- SIL International. (2018). *Technical Notes on FieldWorks Send/Receive*. Retrieved July, 21, 2024, from <https://software.sil.org/fieldworks/wp-content/uploads/sites/38/2018/10/Technical-Notes-on-FieldWorks-Send-Receive.pdf>
- SIL International (2024a). *About SIL*. Retrieved July, 21, 2024, from <https://www.sil.org/about>
- SIL International. (2024b). *Standards*. Retrieved July, 21, 2024, from <https://www.sil.org/language-technology/standards>
- Straka, M., Hajič J., & Straková J. (2016). UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)* (pp. 4290–4297). European Language Resources Association.
- Suchomel, V., & Pomikálek, J. (2012). Efficient Web Crawling for Large Text Corpora. In A. Kilgarriff, & S. Sharoff. (Eds.), *Proceedings of the seventh Web as Corpus Workshop (WAC7)* (pp. 39–43).
- Tiedemann, J. (2016). Finding Alternative Translations in a Large Corpus of Movie Subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 3518–3522). European Language Resources Association.
- Tasovac T., Romary L., Banski P., Bowers J., de Does J., Depuydt K., Erjavec T., Geyken A., Herold A., Hildenbrandt V., Khemakhem M., Lehečka B., Petrović S., Salgado A., & Witt, A. (2024). *TEI Lex-0: A baseline encoding for lexicographic data*. Version 0.9.3. DARIAH Working Group on Lexical Resources. Retrieved July, 21, 2024 from <https://bit.ly/tei-lex-0>
- Text Encoding Initiative (2024). *Getting Started with P5 ODDs*. Retrieved July, 21, 2024, from <https://tei-c.org/guidelines/customization/getting-started-with-p5-odds/>
- Vystrčilová, D. (Ed.) (2002). *Příruční slovník persko-český. Farhang-e hamráh-e fārsí-českí* (Persian-Czech Desk Dictionary). PCHE.

Vystrčilová, D. (Ed.). (2014). *Goftegu.cz. Elektronický tematický česko-perský a persko-český slovník*. Retrieved July, 21, 2024, from <https://goftegu.cz>

Vystrčilová, D. (Ed.). (2017). *Česko-perský slovník. Farhang-e čekí-fársí*. (Czech-Persian Dictionary). Prague.

Vystrčilová, D., Khademi, M., Kříhová, Z., Nachtmann, B., Taucová, R., & Novák, L. (Eds.). (2024a). *Elektronická lexikální databáze indoíránských jazyků: persko-český modul*. Institute of Sociology of the Czech Academy of Sciences, Charles University in Prague. Retrieved July, 21, 2024, from <https://eldi.soc.cas.cz>

Vystrčilová, D., Khademi, M., Kříhová, Z., Nachtmann, B., Taucová, R., & Novák, L. (Eds.). (2024b). *Persko-český mobilní slovník*, 2024-06-20. Institute of Sociology of the Czech Academy of Sciences, Charles University in Prague (forthcoming).

Web Components for TEI Publisher. Version 2.24.1. (2024). Retrieved July, 21, 2024, from <https://github.com/eeditiones/tei-publisher-components>

Wikipedia: The free encyclopedia. (2004). FL: Wikimedia Foundation, Inc. Retrieved July, 21, 2024, from <https://www.wikipedia.org>

XProc 3.0: Specifications. (2024). Retrieved July, 21, 2024, from <https://xproc.org/specifications.html>

Yarshater, E., Daniel, E. L., Ashraf, A., Kasheff, M., & Asthiany, M. (Eds.). (1996–). *Encyclopædia Iranica*. Retrieved July, 21, 2024, from <https://www.iranicaonline.org>

Zook, K. (2015). *FieldWorks 7 XML model*. Retrieved July, 21, 2024, from <https://software.sil.org/fieldworks/wp-content/uploads/sites/38/2016/10/FieldWorks-7-XML-model.pdf>

Acknowledgements

The project ELDI (TL03000369) was co-financed with the state support of the Technology Agency of the Czech Republic as part of the Ěta Program. The development of the Persian web-crawled corpus was, in part, supported by the Slovak VEGA Agency for Science, Project No. 2/0016/21. The contribution of Boris Lehečka in this article in the proceedings was written on the basis of institutional support from the Ministry of Culture for the long-term conceptual development of a research organisation.

The authors thank the anonymous reviewers for their valuable comments, the members of the research team and especially the semantic supervisor Mona Khademi, and Robin Cassling for proofreading this article, the expert consultant in lexicology Michal Škrabal, and the SIL International support team Ann Bush and Ken Zook.

Contact information

Vladimír Benko

Slovak Academy of Sciences, L. Štúr Institute of Linguistics and Comenius University Science Park, UNESCO Chair in Plurilingual and Multicultural Communication
vladimir.benko@juls.savba.sk

Zuzana Kříhová

Charles University, Faculty of Arts
Zuzana.Krihova@ff.cuni.cz

Boris Lehečka

Moravian Library in Brno
lehecka@mzk.cz, boris@daliboris.cz

Darina Vystrčilová

Czech Academy of Sciences, Institute of Sociology
darina.vystrcilova@soc.cas.cz

Cormac Breathnach and Pádraig Ó Mianáin

MAKING A MOLEHILL OUT OF A MOUNTAIN: TECHNICAL AND EDITORIAL CONSIDERATIONS IN PRODUCING THE CONCISE ENGLISH-IRISH DICTIONARY (2020)

A Discussion of Technical and Editorial
Considerations in Producing a 1800-Page Hardback
Dictionary Containing 30,000 Headwords From
an Online Database of 48,000 Headwords

Abstract The *Concise English-Irish Dictionary* (CEID), published in 2020 and the first major English-Irish dictionary published in print form since the 1950s, is a 1800-page hardback dictionary containing 30,000 headwords and 80,000 senses, along with a substantial style and grammar section. Flying in the face of retro-digitisation, this printed dictionary was derived from the *New English-Irish Dictionary* (NEID), a much larger online dictionary published 2013-2017 and containing 48,000 headwords and 145,000 senses. Simply printing the entirety of the online content would have doubled the size of the printed dictionary, so this necessitated a number of measures to whittle the online content down to a single-volume dictionary. This paper outlines some of the challenges and measures involved, such as selection or deselection of lexicographical content, reformatting for print, and the technical process of outputting the same entry to both screen and paper.

Keywords Irish language; online to print conversion; bilingual

1. Introduction

The *Concise English-Irish Dictionary* (CEID), published in 2020, is the first major English-Irish dictionary published since Tomás de Bhaldraithe's *English-Irish Dictionary* in 1959 and the first major bilingual dictionary published in Ireland since Niall Ó Dónaill's *Foclóir Gaeilge-Béarla* [Irish-English Dictionary] in 1977.

It was produced by Foras na Gaeilge, the cross-border implementation body with responsibility for the promotion of the Irish language in Ireland, both North and South, including responsibility for the development of lexicography and terminology. CEID was the final product of the New English-Irish Dictionary project which had started in 2000. The online dictionary which is available free of charge at www.foclóir.ie had been published incrementally from 2013 until 2017, and an accompanying app had been available since 2015. The digital versions contain some 48,000 headwords and 145,000 senses.

The New English-Irish Dictionary project was initially intended to produce a printed dictionary and an accompanying CD-ROM, but by the time the writing phase of the project began in 2008, the rapidly developing digital revolution in lexicography had resulted in online dictionaries replacing printed dictionaries as the primary medium, and ultimately what began as a book and CD-ROM project ended up with the editorial team concentrating on a website and an app.

When the new online dictionary was completed in 2017, it was by no means a given that a printed version would follow. There was a strong argument for turning immediately to the urgent need for a major monolingual Irish dictionary, which would be the first ever of its kind, and a new Irish-English dictionary to succeed the previous one from 1977. However, there were also compelling arguments for producing a printed version: it was always felt that it would do well commercially as it would be the first such English-Irish dictionary in sixty years, and it was also recognised that there was a large demographic of likely Irish-language dictionary users to whom the digital platform was and would remain beyond reach.

A significant challenge resulting from the late decision to produce a printed version was that, as none of the editorial work on the online entries had been done with a print edition of the dictionary in mind, the fact editors had enjoyed zero space limitations up to that point proved to be a major hurdle for the printed dictionary. Another challenge was maintaining one single database for both the digital and printed versions of the dictionary as it was vital to keep both products in the same database to avoid having to amend or correct two versions of each entry continually. The book was finally sent to the printers three years later during the first Covid-19 lockdown in May 2020.¹ In this paper we discuss some of the challenges faced when moving in the unusual direction of online to print, as well as some of our solutions to those problems.

2. Size Matters

The first consideration for the editorial team was the design and physical size of the printed dictionary which would include entries as well as a substantial style and grammar supplement. Early calculations indicated that simply printing the entirety of the online content – all 48,000 headwords, 145,000 senses, almost 130,000 example sentences and around 2 million words – would have run to 3000-3500 pages not including the supplement. That would require two volumes, which would obviously significantly increase costs but would also inevitably greatly reduce the predicted uptake and sales of the product. When the decision was made to go for a one-volume dictionary of around 1500-1800 pages, that meant a 50% reduction of the physical size initially calculated, necessitating editorial measures to select or deselect content from the online version and technical solutions to reformat the dictionary content as space-efficiently as possible.

¹ It was reviewed in the *International Journal of Lexicography* in March 2022 (Pódör, 2022).

3. Trimming by Editing

It was obvious from the beginning of the book project that, aside from a drastic measure such as removing all example sentences, most of the downsizing would have to be done editorially by a staggered selection process, from headword or entire entry level right down to register and domain level. Each level of selection posed its own challenges for the editorial team, as it would prove impossible to achieve a uniform calibration of in/out decision-making among a team of editors, many of whom were working remotely most of the time. While the initial selection at headword level could be based largely on frequency, the lexicographer's dilemma grew as we moved on to sense and example selection where the onus was largely on the individual editor to decide the fate of any given content.

3.1 Selection at Headword Level

The obvious first step was to determine which headwords would be included in the book, and which headwords would not, and to that end the entries in the general NEID dictionary database were divided into three categories for inclusion.

3.1.1 Category 1

Category 1 comprised c. 5,600 headwords selected on frequency basis as outlined in (a) and (b):

(a) The most common 1000 words in English

At the beginning of the translation phase of the NEID project the top 1000 lemmas in order of frequency were given special status and treatment (Ó Mianáin & Convery, 2014, p. 322), and as such they were automatically included in CEID.

(b) Comparison with other dictionaries

Another measure taken during the NEID project was to analyse the headword lists of three medium-sized monolingual English dictionaries and compare them with the headword list for NEID. A score between 1 and 3 was given to each NEID headword depending on its inclusion status in the three control dictionaries. For example, a common headword such as 'table' appeared in all three dictionaries and therefore was given a score of 3. The headword 'bronze' was only present in two dictionaries and was given a score of 2, and so on. Headwords present in all three dictionaries were also included in Category 1 and were automatically included in the headword list for CEID. This amounted to about 4,600 entries.

3.1.2 Category 2: Headwords With a Score of 2 or 1 in the Control Dictionaries

All headwords present in either two or one of the control dictionaries, that weren't in Category 1, were marked as Category 2 and added to the CEID headword list automatically. This amounted to about 6,400 entries.

3.1.3 Category 3: Corpus Frequency Scores and Selection by Editors

The remaining 19,000 or so headwords in CEID were selected using both (a) statistical analysis of the project corpus and (b) coverage checks carried out by editors on a continuous rolling basis.

(a) Corpus frequency scores

A 1.7-billion-word corpus was used as a basis for the English-language content in the NEID project (Ó Mianáin & Convery, 2014). The primary criterion in Category 3 was word frequency in the corpus for the 37,000 headwords not already been included through Categories 1 and 2 above. These headwords were ranked in order of frequency in the corpus and then grouped into 20 bands. There was no set frequency threshold for Category 3 as the final number of headwords in this category depended on how much of the available page-space was being taken up by other categories. Instead, we moved out through the bands of frequency, starting with the top 5%, until we could no longer include any more entries in the book. Approximately 13,000-14,000 headwords were selected using this measure.

(b) Headwords selected by editors

A number of checks were carried out by editors to ensure that CEID would include essential headwords that may not necessarily have been picked up based on frequency in the corpus:

- I. We endeavoured to ensure that specific sets of words such as numbers, directions, colours, musical instruments, etc., were adequately covered and that the main closed sets would be included in their entirety.
- II. Cultural entries relevant to Irish life, culture, history, music and sport, also Hiberno-English.
- III. Encyclopaedic and school-subject entries, of which about 4000 were added, as it was anticipated that the dictionary would probably be used in schools and we were keen to cover a broad range of topics relevant in an educational setting.
- IV. Using Google Analytics, we extracted lists of the most frequently searched items on www.focloir.ie both in the overall seven-year lifespan of the dictionary up to that point and also in the 12-month period leading up to our publication date in 2020, and used those lists to identify additional candidates for inclusion.
- V. Neologisms not in the original DANTE database on which NEID was based but that had been added to the online dictionary since going online in 2013. The online dictionary is regularly updated with new words and phrases that had come to prominence, such as ‘selfie’, ‘craft beer’,

‘emergency accommodation’ and ‘fake news’, and also with words and phrases identified by analysing the unsuccessful searches on the site. A selection of these entries was added to CEID, including a number of Covid-19 terms added in the very final weeks of the book project, such as ‘social distancing’, ‘coronavirus’ and ‘self-isolation’.

- VI. Gender balance, for instance where occupational entries such as ‘doorman’ were reciprocated with ‘doorwoman’ etc.

Approximately 5,000-6,000 headwords were selected using these measures.

The entire headword selection task involving Categories 1-3 resulted in the reduction of the number of headwords from the initial 48,135 entries in NEID to 30,635 in CEID; a 36% reduction.

3.2 Selection at Sense Level

The removal of over a third of the headwords, however, did not by any means equate to a similar reduction in size as the excluded headwords were by nature relatively small in terms of number of senses and examples. In effect, the 36% reduction at headword level translated into an 18% reduction in the number of overall senses and only 8% in the number of usage examples.

The next step in the reduction process was to select or deselect senses under the remaining headwords. Unlike the exclusion process at headword level, which was largely a mechanical process based on word frequency either at corpus level or in the comparison with other dictionaries, the selection of senses for exclusion was primarily down to the judgement of individual editors. Essentially, it was at each editor’s discretion to decide whether or not each sense of the online dictionary entry was sufficiently significant in terms of general language use to merit inclusion in the printed dictionary. In some cases, editors consulted other general bilingual dictionaries in order to determine the importance of word senses in entries. The senior editors would subsequently have an opportunity to promote excluded senses if required, or similarly to demote senses not excluded in the initial editorial pass.

Beginning with an online database of up to 145,000 senses contained in 120,000 sense containers, the final print dictionary was reduced to about 85,000 total senses. As mentioned above, the exclusion of 36% of the headwords accounted for an 18% reduction in sense containers (or about 23% of overall available senses). The exclusion of sense containers by editors in the remaining 30,635 entries led to a further 21% reduction in sense containers (or circa 26% of senses) from the final print product.

3.3 Selection at Micro-Level (Intra-Sense Decisions)

Other items within the remaining sense units were excluded in the CEID entries where they were not considered essential, ranging from examples and their accompanying translations to more granular elements such as domain, style and register labels. Some translations were also excluded, and the sense disambiguators were revised and trimmed where possible. The way usage examples and variants were dealt with is discussed below.

3.3.1 Usage Examples

A lot of the entries in the online dictionary are example-heavy as it is primarily a dictionary for encoding into Irish, but for the printed version the editors were deciding which example sentences would be included. Although there were editorial guidelines for this task, the decisions were largely down to the judgment of individual editors. We recognised from early on that considering every single example sentence would be a far too arduous and time-consuming task and so allowed editors to suppress example sentences on an ad hoc basis.

In the end, over 4000 usage examples in English were suppressed along with their corresponding 4500 translations to Irish. This amounted to a total saving of about 37,500 words, or almost 40 pages of text.

3.3.2 Variant Forms

As a rule, all variants of headwords, phrases and subforms/derived forms were excluded from the print edition of the dictionary. The online version of the dictionary contains almost 10,000 variants of words at various levels. For example, American English spelling of headwords is usually provided (colour/color, analyse/analyze, aesthetic/esthetic). It was felt, however, that the print edition wouldn't suffer for the lack of variants and that their removal would save a lot of space. Indeed, in some cases the variant of a headword is just the headword itself with or without a hyphen (vice captain/vice-captain). This would have looked clunky and unnecessary in the book. This was a relatively straightforward measure to implement but still accounted for a saving of about 21,000 words or over 20 pages of text.

Variants of phrasal verbs were, however, included in the book as they represent a more important part of language usage, for example 'clock off' and 'clock out' are both included.

3.3.3 Variant Forms of Phrases and Idioms

The handling of variants in the case of phrases and idioms was a trickier task. Phrases are by nature quite a bit longer than headwords or phrasal verbs, and as such, the corresponding variant forms are also quite long. We were keen to include the variants deemed linguistically significant, however the inclusion

of ‘to call somebody to account’, ‘to bring somebody to account’ and ‘to hold somebody to account’ would constitute far too much text at the beginning of the phrase. As a solution, it was decided to display a merged version of the phrase, i.e., ‘to call/hold/bring sb to account’. To this end all phrases and their variants were suppressed in the print stylesheet and a new element was introduced that would house the merged version of the phrases.

to call sb to account (*also to bring sb to account, to hold sb to account*)

duine a thabhairt chun cuntais

a iarraidh ar dhuine míniú a thabhairt

she was held to account for the overspend bhí uirthi míniú a thabhairt ar an róchaitheamh

who'll be brought to account for the delay? cé a thabharfar chun cuntais as an moill?

• **to call/hold/bring sb to ~** duine a thabhairt chun cuntais, a iarraidh ar dhuine míniú a thabhairt; **she was held to ~ for the overspend** bhí uirthi míniú a thabhairt ar an róchaitheamh; **who'll be brought to ~ for the delay?** cé a thabharfar chun cuntais as an moill?

Fig. 1: Idiom ‘to call sb to account’ on NEID website (left) and in CEID (right)

3.3.4 Cross-Referencing to Exemplar Entries

The online dictionary has a number of templated sets of entries, where entries such as countries, languages, letters, numbers and pronouns all contain a lot of identical examples to show various usage and grammar elements that can be complicated in the Irish-language equivalents. In the online dictionary, for instance, the entry for ‘you’ contains 14 senses and 156 examples, the entry for ‘Spanish’ contains 4 senses and 18 examples, and the entry for ‘B’ contains 6 senses and 13 examples. As a space-saving measure in the book, however, it was decided to reduce the majority of such entries to their minimum where possible and to cross-reference them to exemplar entries where users could see the usage patterns relevant to the type of word involved. Not all entries could be cross-referenced due to the complexity of the grammar rules in their Irish-language equivalent, as is the case with pronouns and numbers; in the print version, for instance, the entry for ‘you’ has 8 of its original 14 senses and 148 of its original 156 examples.

Ultimately, nearly 500 countries, languages, nationalities and letters of the alphabet were cross-referenced in this manner in the print version, which was a significant saving in terms of number of pages. The printed entry for ‘B’, for instance, was reduced from its original 6 senses and 13 examples to one sense and a cross-reference to the entry for ‘A’, which is a comprehensive entry. Although ‘Spanish’ kept all 4 senses in the print version, all of its 18 examples are suppressed and are replaced by a cross-reference to the entry for ‘German’, an actual reduction in terms of column centimetres from 12 cm to 2.5 cm.

Spanish

1 ADJECTIVE GEOG

Spáinneach *adj1* **C M U**
(de chuid) na Spáinne
he's Spanish is Spáinneach é
a Spanish soldier saighdiúir Spáinneach, saighdiúir de chuid na Spáinne
the Spanish government rialtas na Spáinne

2 ADJECTIVE LING

Spáinnise *gs as adj* **C M U**
a Spanish book leabhar Spáinnise
Spanish class rang Spáinnise
Spanish dictionary foclóir Spáinnise
Spanish exam scrúdú Spáinnise
Spanish teacher múinteoir Spáinnise
a Spanish word focal Spáinnise
the Spanish language an Spáinnis

3 NOUN LING

Spáinnis *fem2* **C M U**
she speaks Spanish tá Spáinnis aici
they're speaking Spanish tá siad ag labhairt Spáinnise, tá siad ag labhairt i Spáinnis
do you know any Spanish? an bhfuil aon Spáinnis agat?
how do you say 'Ireland' in Spanish? cén chaoi a ndeirtear 'Éire' i Spáinnis?
what's the Spanish for 'food'? cad é an Spáinnis ar 'bia'?
say that in Spanish abair é sin i Spáinnis
a film in Spanish scannán Spáinnise
a degree in Spanish céim sa Spáinnis

the Spanish *NOUN GEOG*

na Spáinnigh *mpl1* **C M U**
muintir na Spáinne

Fig. 2: The entry 'Spanish' on the NEID website

Spanish **ADJ** ① GEOG Spáinneach
a1, (de chuid) na Spáinne ② LING
Spáinnise *nmod*
N LING Spáinnis f2
the Spanish **N** GEOG na Spáinnigh
mpl1, muintir na Spáinne ▶ for
usage see German

Fig. 3: The entry 'Spanish' in CEID

3.3.5 Domain Labels and Sense Disambiguators

On the NEID website, the domain labels and sense disambiguators are presented on the first line of each sense, as in the first two senses of the entry 'access rights' below. In the case of the printed dictionary, however, it was deemed superfluous to include both the domain label and the sense disambiguator unless absolutely necessary.

access rights

- 1 *NOUN LAW* to see child
cearta rochtana 4» C M U
cead cuairte *familiar* 4» C M U
he wants access rights to his children teastaíonn cead cuairte lena pháistí uaidh
- 2 *NOUN COMP* to view file etc
cearta rochtana 4» C M U
- 3 *NOUN* to enter location
cearta dul isteach

Fig. 4: Entry 'access rights' on NEID website

Running a job in DPS to *suppress all domain labels* or to *suppress all disambiguators* in the print style sheet was too risky considering the amount of data involved and would have necessitated a large checking task afterwards. Instead, we created a spreadsheet of all the senses in the book that contained both domain labels and sense disambiguators. The ideal scenario for us was to use the domain labels as often as possible as they are much shorter in length than the disambiguators. We went through the spreadsheet – some 8,500 lines – and decided on whether to suppress the domain label or the disambiguator. In a small number of cases, we kept both. Using the spreadsheet, a job was run on the data in DPS and the @hide=print attribute (discussed in more detail below) was added to the selected elements. The example from the spreadsheet below shows the decisions in column four on which element – domain label or disambiguator – would be suppressed in the book.

Headword	Domain label	Sense disambiguator	HIDE?
A	mus	note, key	sense dis.
A	med	blood group	domain
abbreviation	ling	of word, phrase	sense dis.
abdicate	pol	renounce throne	sense dis.
abdication	pol	renouncing of throne	sense dis.
aberrant	biol	of gene, cell, etc	sense dis.
abnormal	stat	of curve etc	domain
abolitionist	law, pol	supporting abolition of sth	sense dis.
abolitionist	pol, hist	supporting abolition of slavery	sense dis.
abort	comp, aero	stop prematurely	sense dis.
abrasion	med	injury	sense dis.
abridge	law	limit	sense dis.
abrogate	gov, law	repeal	sense dis.
abscess	med	infected swelling	sense dis.
abscond	police	escape	domain
absence	bus-adm, ed	being away from somewhere	sense dis.
absent	sociol	not part of household	domain
absentee	hist	not living in certain place	domain
absolution	rel	forgiveness of sin	sense dis.
abstract	art	of artwork, artist	sense dis.
abstract	art	artwork	sense dis.
accent	ling	long accent in Irish	domain
accent	ling	diacritic	domain
accent	ling	stress	domain
accent	ling, mus	stress	sense dis.
access	sociol	for disabled	domain
access	telecom	connection	domain
access	law	contact with child	domain
access	telecom	connect with	domain

Fig. 5: Example from domain and disambiguator spreadsheet

4. Technical Measures

All of Foras na Gaeilge's dictionary projects are hosted in IDM's Dictionary Production System (DPS), and all technical solutions involve customisations within DPS.

4.1 Suppressing not Deleting: the @hide=print Attribute

As mentioned in the introduction, both the digital and printed dictionaries were being produced from the one dictionary database as we didn't want to have two separate databases to maintain – one for the book and one for the website. Consequently, it was of vital importance that dictionary content within the database would not be deleted in order to save space in the book. There is a huge amount of content on display on the website that could not be accommodated in the book. It was decided, therefore, that content that was to be excluded from the print dictionary would be suppressed or 'hidden' in the stylesheet rather than deleted from the database entirely. The @hide attribute was added to most elements in the database which could be set to any of the following values: never, always, print, or online.

The @hide=print attribute gave editors the option of hiding a sense, a string of text, a label or some other element in the database from the print stylesheet, without having to delete the element altogether and thus losing it from the online version of the website. As well as a WYSIWYG stylesheet to see the entry in the print layout, a previewer was also developed within DPS which would 'grey out' material that was set to @hide=print. In the example below, the fourth sense is fully hidden or suppressed as well as some example sentences in sense one and the sense disambiguator in sense three.

addition
1 *NOUN* sb or sth added
 it's a great addition to the village cuireann sé go mór leis an sráidbhaile
 she was a last-minute addition to the list cuireadh ar an liosta ag an nóiméad deiridh í
 Seán is the latest addition to the family is é Seán an páiste is déanaí a rugadh sa chlann
 an index would be a useful addition bheadh sé áisiúil dá mbeadh innéacs leis
 any amendments or additions? aon rud le leasú nó le cur leis?
 to make an addition to sth breis a chur le rud, cur le rud
 they made some additions to the text chuir siad roinnt rudaí leis an téacs

2 *NOUN* process of adding
 since the addition of law to the programme ó cuireadh an dlí leis an gclár
 it's sweetened by the addition of fruit juice milsítear é trí shú torthaí a chur leis

3 *NOUN MATH* of numbers
 suim iú *m asc*
 suim iúchán *m asc 1*
 repeated addition s uimíú leantach

3 *NOUN (US) CONSTR* extension
 píos a breise
 síneadh *m asc 1*
 to build an addition to a house síneadh a chur le teach

Fig. 6: Entry 'addition' in the hybrid CEID stylesheet displaying content excluded from the print dictionary

4.2 Trimming by Formatting

Once the lexicographical content of the printed dictionary had been decided upon, the next challenge was to find a workable way of presenting that content in as concise a manner as possible while ensuring that the dictionary was user-friendly. Quite a lot of time went into typographic design features (fonts, colours, spacing, leading, etc) and other aesthetic elements. We used a sans-serif font for the English-language content (Atlas Grotesk Bold) in 7.8pt and a serif font for the Irish-language content (Lyon Text Regular No. 2) in 8.5pt. This font is larger than that which would usually be expected in a dictionary of this size, however as this was the first major bilingual Irish dictionary in over 40 years, we were keen to produce a premium quality book that was both readable and functional.

The layout of entries in the online dictionary is quite vertical in style, with a line break before every sense, example and base translation, for example:

gentleman

1 *NOUN* man

fear *masc1* 🗣️ **C M U**

ladies and gentlemen! *a dhaoine uaisle!*

an older gentleman *seanfhear* *masc1*, *seanduine* *masc4*

he's a gentleman of leisure *tá saol an mhadra bháin aige, níl cúram sa saol air*

2 *NOUN* polite, considerate man

fear uasal 🗣️ **C M U**

he's a true gentleman *fear uasal cruthanta is ea é, corp an duine uasail atá ann*

Fig. 7: Entry 'gentleman' on NEID website showing use of line breaks

For the printed version, we needed to remove as much white space as possible and to fit the entries into as small a space as possible without detriment to the readability or usability of the dictionary.

4.2.1 The Point of no RETURNS

Keeping line breaks to a minimum was the first priority in the print edition. Each new entry would, of course, have to start with a line break but all other options were on the table. Having consulted a range of similar bilingual dictionaries and considered all the options in InDesign – the publishing software used to design the book – it was decided to insert line breaks based on the following criteria, something we felt was a happy medium between readability and space-efficiency:

- before each new entry;
- before each part of speech, except the first one in the entry;
- before each phrasal verb;
- before each multi-word entry block (i.e., containing phrases and/or phrasal verbs);
- before each phrase and phrasal verb within those blocks, except the first one;
- before each subform of entries.

Apart from the instances mentioned above, content would run on from line to line. The most discussed item here was the decision on whether to put a line break between part of speech blocks, something that works reasonably well in mid-sized and large entries but looks somewhat clunky or untidy in small entries with no example sentences. In ‘abseil’ below, for example, you can see from the surrounding entries that there would have been space on the first line for the POS marker and for the first word of the next sense. On the other hand, in the mid-sized entry ‘acid’ the POS marker starting on a new line makes the entry clearer and more readable. This becomes even more apparent in large entries.

abscond v (*escape*) v_I éalaigh, teith
abseil v_T & v_I rópáil anuas
 N rópáil anuas
abseiling N rópadóireacht anuas
absence N ① (*non-existence*) easpa
f4, uireasa *f4*, ceal *m4*; in the ~ of

Fig. 8: Entry ‘abseil’ in CEID

acid N ① CHEM, MED aigéad *m1*;
 folic ~ aigéad fólach; lactic ~ aigéad
 lachtach; stomach ~ aigéad goile ②
 (*drug*) aigéad *m1*; an ~ trip siabhrán
 aigéid
 ADJ ① CHEM aigéadach *a1*, aigéid
nmod; ~ soil ithir aigéadach ② (*crit-*
ical) binbeach *a1*, nimhneach *a1*; he
 gave me an ~ look thug sé féach-
 aint bhinbeach orm, thug sé amharc
 nimhe orm ③ (*sour*) searbh *adj*, géar
a1; an ~ flavour blas géar

Fig. 9: Entry ‘acid’ in CEID

4.2.2 Swung Dashes

It is a common practice in printed dictionaries to use swung dashes or tildes to denote the headword within that headword’s entry (Bringinghurst, 2012, p. 318), however it wasn’t as simple as replacing every instance of the headword with a swung dash and moving on.

It was a simple decision to use swung dashes in example sentences. We felt users could easily understand the convention and also be able to handle ‘~s’ or ‘~ed’ for plurals or verbs.

adjust v ① (*alter slightly*) vT socraigh, deisigh; **he ~ed his tie** shocraigh sé a charbhat, dheisigh sé a charbhat; **you can ~ the acidity of the soil** is féidir aigéadacht na cré a shocrú ② (*for accuracy, effect, etc*) vT

Fig. 10: Part of entry 'adjust' in CEID

We decided to also use the swung dash in the headings of phrases and idioms. They were generally in close proximity to the headwords and, again, we felt users wouldn't have a problem dealing with them. See example below from 'addition':

PHRASES ▪ **in ~** chomh maith leis sin, freisin, fosta, ina theannta sin, lena chois sin; **in ~, it's available online** chomh maith leis sin, tá sé le fáil ar líne; tá sé le fáil ar líne chomh maith
▪ **in ~ to** PREP chomh maith le, i dteannta, le chois; **in ~ to music, they do art** chomh maith le ceol, déanann siad ealaín

Fig. 11: Phrases and idioms in entry 'addition' in CEID

One complication that is evident in the examples under 'addition,' however, is that the swung dash in the example sentences only stands for the headword of the entry rather than the sub-headword or the phrase in question. This is also an issue in the case of variant forms of phrasal verbs:

act on (*also ~ upon*) ① (*take action*) vT
gníomhaigh de réir, déan de réir;
to ~ on the recommendations of sb gníomhú de réir moltaí duine;

Fig. 12: Phrasal verb 'act on' in CEID

In the case above, an unskilled user could take it that the swung dash stands for the entire phrasal verb and as a result could think that the variant form of 'act on' in this case is 'act on upon' or even that the example sentence could be read as 'to act on on the recommendations of sb'. On balance, however, it was felt that in general users of the dictionary would understand what the swung dashes stand for in such cases. Secondly, it was not guaranteed that a swung dash in place of the full phrasal

verb would be any clearer than the approach we employed. And finally, the value of the amount of space saved with the swung dashes was too high to forego in order to include full headwords in the phrasal verb sections.

4.2.3 Abbreviations

In the online version of NEID, the parts of speech are spelled out in full for the English headwords and partially for the Irish translations, however they were all abbreviated further in the print edition and explained in the front matter of the dictionary in English and in Irish. The resultant space-saving may be observed in the entry ‘affray’ below, where *noun* became *n*, and *masc1* [masculine noun, first declension] became *m1*.

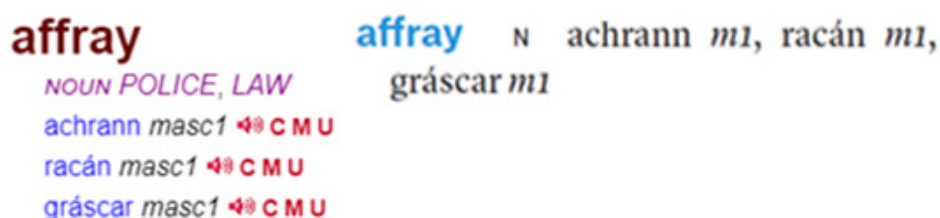


Fig. 13: Entry ‘affray’ on NEID website (left) and in CEID (right)

5. Conclusion

The measures outlined above reduced the dictionary content of CEID from an initial estimate of between 3,500 and 4,000 pages to 1,700 pages of dictionary content plus 100 pages of front matter and supplementary material. The dictionary was printed in 2020 and reprinted in 2021.

References

- Bringhurst, R. (2012). *The Elements of Typographic Style* (4th ed.). Hartley & Marks Publishers.
- de Bhaldraithe, T. (Ed.). (1959) *English-Irish Dictionary*. An Gúm.
- Ó Dónaill, N. (Ed.). (1977). *Foclóir Gaeilge-Béarla [Irish-English Dictionary]*. An Gúm.
- Ó Mianáin, P. (Ed.). (2020). *Concise English-Irish Dictionary*. Foras na Gaeilge.
- Ó Mianáin, P. (Ed.). Foras na Gaeilge. (2024, January). Retrieved from New English-Irish Dictionary: <https://www.focloir.ie>
- Ó Mianáin, P., & Convery, C. (2014). From DANTE to Dictionary: The New English-Irish Dictionary. In A. Abel, Ch. Vettori, & N. Ralli (Eds.), *Proceedings of the 16th EURALEX International Congress* (pp. 807–817). Institute for Specialised Communication and Multilingualism.

Ó Mianáin, P., & Convery, C. (2014). New English-Irish Dictionary. *Journal of the Dictionary Society of North America*, 35, 318–333.

Pődör, D. (2022). Pádraig Ó Mianáin (chief editor). 2020. Concise English-Irish Dictionary. Foclóir Béarla-Gaeilge. *International Journal of Lexicography*, 35(1), 129–135.

Contact Information

Cormac Breathnach

Foras na Gaeilge
cbreathnach@forasnagaeilge.ie

Pádraig Ó Mianáin

Foras na Gaeilge
pomianain@forasnagaeilge.ie

Elina Chadjipapa and Zoe Gavriilidou

HELIX: A BILINGUAL ILLUSTRATED DICTIONARY FOR GREEK HERITAGE LEARNERS

Abstract The recent development of the *Curriculum for Teaching Greek as a Heritage Language: A Framework for Teachers* (Gavriilidou & Mitsiaki, 2022) underscored the need for a dictionary to serve as supplementary material during the curriculum's implementation at Greek Community schools in the USA. This presentation aims to introduce the Greek Heritage Language Learners' Illustrated Lexicon (Helix), an online, bilingual, illustrated dictionary designed for Greek Heritage Learners aged 5-10 years living in USA, Canada, Australia, etc. Helix is intended to enhance the teaching resources available to educators who teach Greek as a heritage language. It provides word definitions based on the linguistic competence of heritage speakers of Greek, facilitating efficient vocabulary acquisition and the development of students' language skills. Specifically, it aims to serve as the primary dictionary for Greek Heritage students, whether used in the classroom or outside, to practice their reading, writing, speaking, and listening skills.

Keywords Greek heritage speakers; heritage learners; dictionary making; pedagogical lexicography; lexicographic protocol

1. Introduction

A heritage language speaker (HS) is an early bilingual who typically acquires their parents' language at home. A typical HS profile includes a child growing up in a family where the home language differs from the host country's language. However, as an HS grows and attends school, proficiency in the majority language often overshadows the heritage language, limiting its use within the ethnolinguistic community or at home. Despite the home language being used by parents or caretakers, many young HSs prefer to communicate in the dominant language (Polinsky & Kagan, 2007; Gavriilidou & Mitits, 2021).

This linguistic shift leads to significant vocabulary gaps, hindering comprehension and expression in the heritage language, both written and oral (Gavriilidou & Mitits, 2021). Additionally, when the alphabet is non-Latin, as in the case of Chinese, Korean, and Greek heritage speakers (HSs) in USA or Canada who initially became literate using the Latin alphabet but must also learn a non-Latin alphabet (such as Greek in our case) to become literate in their heritage language, reading and writing become more challenging (Polinsky & Kagan, 2007).

Most research on heritage languages has focused on speakers whose home languages are Spanish, Chinese, Russian, and Polish (see Carreira & Kagan, 2011). Greek Heritage Speakers (GHSs) have only recently become a subject of study (Gavriilidou & Mitits, 2019; 2021). The Greek Heritage Language Corpus (GHLC) (Gavriilidou et

al., 2019), developed as part of the project “Varieties of Greek as a Heritage Language: Creation of a Corpus and Comparative Study” (MIS 5006199, Scientific Coordinator Professor Zoe Gavriilidou), is the first spoken corpus of Greek as a heritage language. It includes data from first, second, and third-generation Greek heritage speakers living in Chicago, Moscow, and Saint Petersburg. The GHLC has been instrumental in profiling this group of heritage speakers. Its collection and compilation aim to:

- Improve theories of heritage language acquisition, focusing on lexical, morphosyntactic, and pragmatic features of Greek heritage language varieties.
- Contribute to the design of needs-analysis based tools and educational resources and materials, such as the Helix dictionary, and to the selection and structuring of teaching content (syllabi design), vocabulary teaching (false cognates, blends, etc.), and grammar teaching (prioritization of grammatical structures based on heritage learner needs).

The content of the GHLC reflects its purpose and significantly supports research in Heritage Greek.

Teaching heritage languages is widely recognized as highly important today. Heritage speakers (HSs) typically share a common profile but vary significantly in their receptive and productive command of the heritage language (Benmamoun et al., 2013). Teaching heritage learners (HLs) is challenging due to their distinct educational needs (Gavriilidou et al., 2023). Teachers must engage students by selecting topics with practical significance to prevent them from disengaging from the learning process or failing to regain proficiency in the home language (Wiley, 2008).

To address these challenges, the new *Curriculum for Teaching Greek as a Heritage Language: A Framework for Teachers* (Gavriilidou & Mitsiaki, 2022), compiled to be used in Greek Community daily or Saturday Schools in the US, employs both macro- and micro-level approaches aligned with the principles of the American Council on the Teaching of Foreign Languages (ACTFL). The curriculum strategically integrates dictionary activities as part of a spiral approach, enhancing vocabulary acquisition and fostering overall language growth across various competence levels for heritage language learners. However, this curriculum needs to be complemented with reference tools such as dictionaries and lesson plans to ensure successful classroom implementation. This is precisely the purpose behind creating Helix.

2. Dictionary Use

Research on dictionary use and its impact on language skills and vocabulary learning has long engaged scholars and remains a significant area of interest. The results of the research highlight the importance of using dictionaries during the learning process, both in terms of comprehension (Wingate, 2002; Shieh & Freiermuth, 2010; Shen, 2013, Hamdi, 2015) and production (Chun, 2004; Elola et al., 2008; Wolfer et al., 2016), as well as in vocabulary learning (Chen, 2011; Ma & Cheon, 2018). Consequently, on one hand,

the use of dictionaries should be part of the educational process and integrated into it, and on the other hand, students should be effective dictionary users (Chadjipapa et al., 2020). This necessitates training students in dictionary use, which can be achieved through explicit instruction in dictionary use strategies (Gavriilidou et al., 2024). The prerequisite for all this is the existence of appropriate and well-designed dictionaries that meet the specific learning conditions and needs of the users (Tono, 2001).

Recognizing the pedagogical benefits of dictionary use in vocabulary learning and on reading and writing, the curriculum strategically integrates dictionary activities as part of a spiral approach, enhancing vocabulary acquisition and fostering overall language growth across various competence levels for heritage language learners.

The creation of the Greek Heritage Language Learners' Illustrated Lexicon (Helix) is an essential complement to the new curriculum. Helix innovates in several aspects:

- **Grounded in the Needs of Specific Learners:** It is tailored to meet the unique requirements of Greek Heritage Language learners.
- **Curriculum Integration:** It is the first dictionary designed specifically to support the new *Curriculum for Teaching Greek as a Heritage Language* in the USA.
- **Pedagogical Focus:** This dictionary addresses the educational needs of both heritage learners and teachers.
- **Illustrated and Interactive:** By providing images and animations, it assists heritage learners who may struggle with reading in their home language.
- **Accessible and User-Friendly:** Helix is an electronic dictionary freely accessible on the Internet, prioritizing flexibility and ease of use.

3. Helix: Steps and Design Principles

Helix is an online, bilingual, illustrated dictionary designed for Greek Heritage Learners aged 5–10 years. It encompasses the proposed vocabulary outlined in the syllabi for Kindergarten (Pre-K) and Novice (K1-K3) levels included in the *Curriculum for Teaching Greek as a Heritage Language: A Framework for Teachers*. Due to the language variance that a heritage speaker (HS) can demonstrate (Polinsky, 2015), vocabulary from earlier stages (Pre-K1) was also included in the dictionary. This inclusion ensures comprehensive coverage and addresses the diverse linguistic proficiencies of heritage learners.

In order to ensure a transparent and reproducible process of its compilation (Gavriilidou & Garoufos, 2022), Helix adopts the six phases of the lexicographic protocol proposed in Klosa (2013), for the creation of electronic dictionaries:

1. **The phase of preparation:** The preparation phase encompasses essential lexicographical processing and research necessary for crafting a dictionary.

It's a pivotal stage where crucial decisions are made regarding the dictionary's requirements, workflow, time allocation for completion, and cost estimation. Throughout this process, careful consideration is given to factors such as the type of dictionary, its intended users, size, and structure (Schierholz, 2015). In this phase, it was decided that, based on the language criterion, Helix would be bilingual (Greek-English) dictionary. Its coverage would be specialized addressing the needs of a specific audience; It is a school dictionary for Greek Heritage Speakers. Finally, it was decided to have a small size (see macrostructure below) and to be organized in three different ways (see below).

2. The phase of data acquisition: This phase follows the completion and finalization of the dictionary preparation phase and encompasses the data collection process (see 3.1 below).
3. The phase of computerization: The text collection process was followed by the phase of digitization, which involved organizing the content into files suitable for processing (.txt) by text analysis software.
4. The data processing phase: During this stage, we determine the final list of lemmas to be included in the dictionary.
5. The phase of data analysis: The data analysis phase is a pivotal stage in which the macrostructure and microstructure of the dictionary, along with any other pertinent information related to the lemma entries, are thoroughly examined and elucidated and
6. The phase of preparation for online release: Helix was designed and implemented as a web application.

3.1 Macrostructure

To determine Helix's macrostructure, we relied on several key elements:

a) Curriculum Vocabulary: The vocabulary suggested by the new *Curriculum for Teaching Greek as a Heritage Language* for each thematic unit in the syllabi for PreK and K1-3 was a primary source. This alignment ensures that Helix supports and enhances the curriculum's implementation.

b) Specially Compiled Corpus: A dedicated corpus, consisting of age-appropriate texts, was specifically compiled for this dictionary. This corpus includes texts from the databank of the syllabi of the official exams from the Centre for the Greek Language (CGL) for the certification of Greek language proficiency concerning level A1 (period 2000–2007). It provided a realistic and relevant lexical database tailored to the needs of young Greek heritage learners.

c) Reference to Similar Dictionaries: The headword lists from similar illustrated dictionaries in other languages, such as the Wordsmyth Illustrated Learner's

Dictionary (WILD), were consulted. This comparative analysis helped to refine the selection of headwords, ensuring that Helix meets high pedagogical standards and offers a robust learning tool.

Initially, we began with a selection of 700 lemmas based on the vocabulary outlined in the new *Curriculum for Teaching Greek as a Heritage Language: A Framework for Teachers*. This curriculum served as our primary guide for determining which words to include in Helix.

Following this initial selection, we organized the lemmas into categories based on word families to ensure a structured and cohesive approach to vocabulary presentation. However, the process didn't stop there.

Further refinement of the lemmas was achieved through the analysis of a corpus specific to this project and by consulting similar dictionaries. The corpus analysis provided insights into the frequency and usage patterns of certain lemmas, helping us to prioritize and adjust our selection.

In some cases, the corpus analysis revealed that certain lemmas were highly frequent in usage but were not included in the main proposed vocabulary. Additionally, these lemmas were found to be relevant to thematic fields covered in higher educational levels.

For example, the word “ψωμί” (‘bread’) was not initially included in the proposed vocabulary for PreK or K1-K3 educational levels. However, through corpus analysis, it was discovered that this word had high frequency and relevance, particularly in Intermediate Low K4 & K5 levels. Students might encounter and use this word naturally during the proposed activities outlined in the syllabus.

Therefore, we made the decision to include such lemmas in the macrostructure of Helix. This ensures that the dictionary not only aligns with the proposed vocabulary but also anticipates and accommodates the evolving linguistic needs of heritage language learners across different educational levels.

Through this procedure, we meticulously refined our selection, resulting in a comprehensive list of 1500 lemmas.

The dictionary entries are arranged in three different ways according to:

- a) the alphabetic order of the entries;
- b) the thematic field that they belong. Thus, the entries can be searched and found according to word families (animal, people, months, plants, etc.);
- c) to the part of speech (noun, verb, adjective).

The environment of Helix can switch from English to Greek.

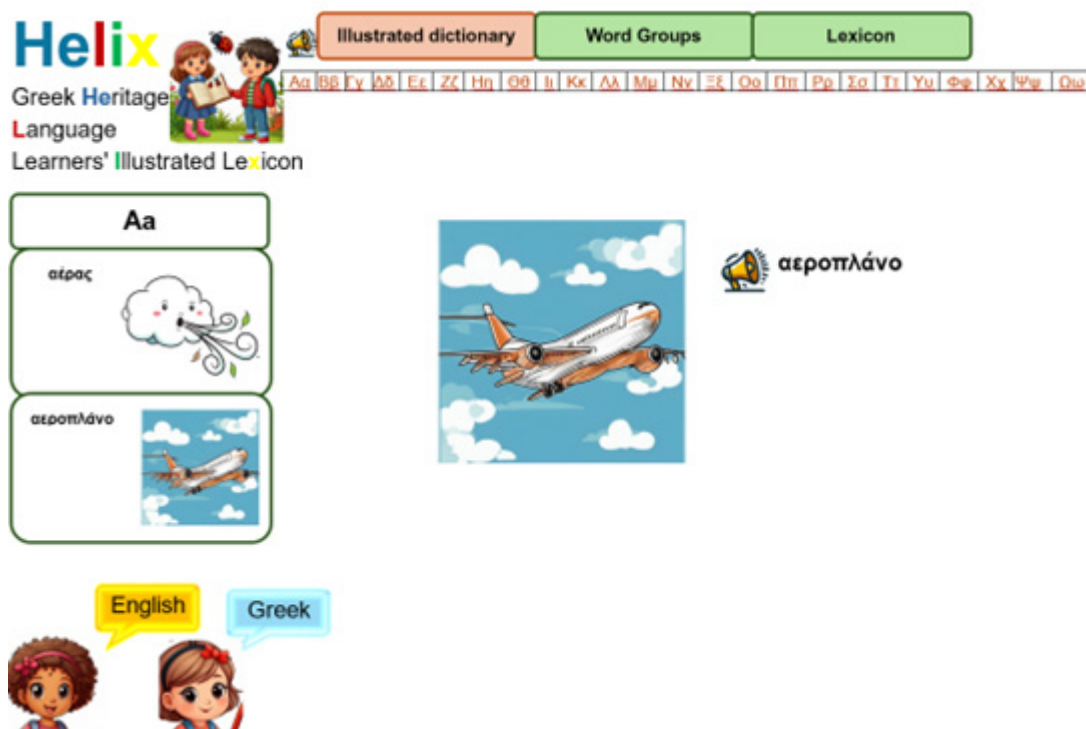


Fig. 1: Helix, Alphabetical order

The word families/groups that lemmas have been categorized and can be found in the field “Word Groups” are:

- animals (mammals, birds, fish, insects, reptiles and amphibians, young animals and animal parts);
- plants (flowers, trees, parts of a plant, tree parts);
- people (body, feelings, description, nationality, family members, professions);
- months, days and weather;
- food (fruit, vegetables, meat and seafood, breakfast, dinner, lunch, taste);
- music;
- transportation;
- home (furniture, clothes, places at home, things in a room, things happen at home);
- classroom (furniture, things in a classroom, people at school, courses);
- toys;
- numbers, shapes, colours;
- anniversaries (Christmas, Easter, 25 of March, 28 of October).

3.2 Microstructure

In the “Lexicon” section, depicted in Figure 2, detailed information about the orthographic form and semantics of the lemmas is provided. This includes definitions, examples, synonyms, antonyms, and word families. Additionally, pragmatic labels are incorporated to enhance user understanding, offering cultural insights and contextual connections to the broader world. Most of the semantic information appears in a second window when the user selects the “**see more**” option beneath the photo. This design choice accommodates students of various educational levels.



Fig. 2: Semantic details, meanings

Furthermore, it was decided that each entry would include details about the phonetic forms of lemmas in various arrangements (alphabetical and thematic). Users can listen to the pronunciation of each entry by clicking on the corresponding button. The words were recorded by the researcher in the +MorPhoSe/SynMorPhoSe Linguistics Laboratory using specialized equipment. This feature aims to provide users with easy access to the correct pronunciation of all the words.

In selecting illustrations, particular attention was paid to the age and cultural background of the users (Biesaga, 2016). Many of the images were generated using Image Creators equipped with Artificial Intelligence capabilities. However, it's important to note that these image generators typically offer only a limited number of free attempts for image creation. Consequently, the assistance of a graphic artist was necessary to ensure the completion of the dictionary.

4. Conclusion

As of now, most phases of the lexicographical process involved in creating Helix have been completed, including preparation, data acquisition, computerization, and data processing. Upon the completion of all dictionary construction phases, feedback will be solicited from target users to assess its user-friendliness, usability, and quality (de Schryver & Prinsloo, 2000; Klosa, 2013).

Once Helix is finalized, intervention programs related to dictionary use will be implemented in Greek Heritage Language (GHL) classrooms. This initiative serves a dual purpose: a) to enhance heritage language speakers' strategic dictionary use and effectiveness, and b) to evaluate the dictionary's utility as an educational tool for educators in heritage language classrooms (Polinsky & Kagan, 2007).

The insights gained from these initiatives will inform decisions regarding enriching the dictionary to cater to intermediate low (K4-K5) and intermediate mid-high (K6-K8) heritage language speakers.

This study offers invaluable insights into pedagogical lexicography, highlighting the strategic integration of learner's dictionaries with existing curricula and educational materials. This synergistic approach aims to establish a comprehensive framework for vocabulary development, thereby significantly contributing to the enhancement of language proficiency.

References

- Benmamoun, E., Montrul, S., & Polinsky, M. (2013). Heritage languages and their speakers: opportunities and challenges for linguistics. *Theoretical Linguistics*, 39(3-4), 129-181.
- Biesaga, M. (2016). Pictorial illustration in dictionaries. The state of theoretical art. In T. Margalitadze, & G. Meladze (Eds.), *Proceedings of the XVII EURALEX International Congress* (pp. 99-108). Ivane Javakhishvili Tbilisi University Press.
- Carreira, M., & Kagan, O. (2011). The results of the national heritage language survey: Implications for teaching, curriculum design, and professional development. *Foreign Language Annals*, 44(1), 40-64. <https://doi.org/10.1111/j.1944-9720.2010.01118.x>
- Chadjipapa, E., Gavriilidou, Z., Markos, A., & Mylonopoulos, A. (2020). The Effect of Gender and Educational Level on Dictionary Use Strategies Adopted by Upper-elementary and Lower-secondary Students Attending Greek Schools. *International Journal of Lexicography*, 33(4), 443-462.
- de Schryver, G.-M., & Prinsloo, D. J. (2000). Dictionary-Making Process with 'Simultaneous Feedback' from the Target Users to the Compilers. In U. Heid, S. Evert, E. Lehmann, & C. Rohrer (Eds.), *Proceedings of the Ninth EURALEX International Congress* (pp. 197-209). Inst. für Maschinelle Sprachverarbeitung.

- Gavriilidou, Z., & Garoufos, A. (2022). The lexicographic protocol of MikaelaLex: a free online school dictionary of Greek accessible for visually-impaired senior elementary children. In A. Kloosa-Kückelhaus, S. Engelberg, C. Möhrs, & P. Storjohann (Eds.), *Dictionaries and Society*, (pp. 563–568). IDS-Verlag.
- Gavriilidou, Z., Markos, A., & Konstantinidou, E. (2024). How Can We Raise Strategic Dictionary Use in the Classroom: The Effect of a Dictionary Awareness Program on Dictionary Use Strategies. *Lexikos*, 34(1), 99–122. <https://doi.org/10.5788/34-1-1882>
- Gavriilidou, Z., & Mitits, L. (2019). Profiling Greek heritage language speakers in the USA and Russia. *European Journal of Language Studies*, 6(1), 28–42.
- Gavriilidou, Z., Mitis, L., Mavrommatidou, S., Chatzipapa, E., & Dourou, C. (2019). The compilation of Greek heritage language corpus (GHLC): a language resource for spoken Greek by Greek communities in the U.S. and Russia. *European Journal of Language Studies*, 6(1), 61–74.
- Gavriilidou, Z., & Mitits, L. (2021). The Socio-linguistic Profiles, Identities, and Educational Needs of Greek Heritage Language Speakers in Chicago. *Journal of Language and Education*, 7(1), 80–97. <https://doi.org/10.17323/jle.2021.11959>
- Gavriilidou, Z., & Mitsiaki, M. (2022). *Curriculum for teaching Greek as a Heritage Language: a framework for teachers*. Ed. Office of Greek Education of the Greek Orthodox Archdiocese of America. 2kProject.
- Gavriilidou, Z., Mitsiaki, M., & Mitits, L. (2023). Strengthening Heritage Greek Vitality in the United States through a Community-Based Curriculum. *Heritage Language Journal*, 20(1), 1–28. <https://doi.org/10.1163/15507076-bja10021>
- Hamdi, C. (2015). The Effects of Electronic Dictionary Use on Reading Comprehension and Vocabulary Retention of EFL Students. *Arab World English Journal*, 180–191.
- Klosa, A., (2013). The lexicographical process (with special focus on online dictionaries). In R. H. Gouws, U. Heid, W. Schweickard, & H. Wiegand (Eds.), *Dictionaries. An International Encyclopedia of Lexicography: Supplementary Volume. Recent developments with focus on electronic and computational lexicography* (pp. 517–524). Walter de Gruyter.
- Ma, J. H., & Cheon, H. J. (2018). An experimental study of dictionary use on vocabulary learning and reading comprehension in different task conditions. *International Journal of Lexicography*, 31(1), 29–52.
- Polinsky, M., & Kagan, O. (2007). Heritage languages: In the ‘wild’ and in the classroom. *Language and Linguistics Compass*, 1, 368–395.
- Polinsky, M. (2015). Heritage languages and their speakers: State of the field, challenges, perspectives for future work, and methodologies. *Zeitschrift fuer Fremdsprachwissenschaft*, 26, 7–27.
- Shen, Z. (2013). The effects of vocabulary knowledge and dictionary use on EFL reading performance. *Canadian Center of Science and Education*, 6(6), 77–85.

Shieh, W., & Freiermuth, M. R. (2010). Using the DASH Method to Measure Reading Comprehension. *TESOL Quarterly*, 44(1), 110–128.

Tono, Y. (2001). *Research on Dictionary Use in the Context of Foreign Language Learning: Focus on Reading Comprehension*. Lexicographica Series Maior 106. Niemeyer.

Wiley, T. G. (2008). Dialect speakers as heritage language learners: A Chinese case study. In D. Brinton, & O. Kagan (Eds.), *Heritage Language: A New Field Emerging* (pp. 91–106). Routledge.

Contact information

Elina Chadjipapa

Democritus University of Thrace, Greece
elinaxp@hotmail.com

Zoe Gavriilidou

Democritus University of Thrace, Greece
zoegab@otenet.gr

Rajna Dragičević, Yury Makarov, Daria Ryzhova,
Yulia Shapich, and Ekaterina Yakushkina

A NEW SERBIAN-RUSSIAN DICTIONARY

Abstract The paper presents a New Serbian-Russian dictionary and the main principles of its development. We use the most recent explanatory dictionary of Serbian, published by the Serbian Academy of Sciences and Arts in 2018, as a starting point. However, we refine both the word list and the entry structure to meet the requirements of a bilingual edition. We consult text corpora of modern Serbian to check frequencies and define contexts of word usage. The paper describes the criteria for word list formation, discusses the problem of lexical equivalence in closely related languages, and articulates our guidelines for representing polysemy. The dictionary will be published in both traditional print format and as a website. The digital version will be implemented using the OnLex platform; we detail its functionality and highlight the new possibilities it offers to editors and potential users.

Keywords Serbian; Russian; dictionary; online lexicography; OnLex

1. Introduction

The number of various electronic resources, such as corpora, dictionaries, and electronic databases for various languages, including Slavic, is continuously and rapidly growing. Despite this, there is currently no sufficiently complete and up-to-date dictionary which would be capable of fulfilling the needs of modern language users interested in Serbian and Russian. The most recent bilingual dictionary was published in 1957 by Iliya Tolstoy (expanded edition — 1970) (Tolstoy, 1970). Since then, not only have the vocabularies of both languages changed but also the approaches to writing and publishing dictionaries as well as lexical analysis in general.

To fill in this gap, we are creating a New Serbian-Russian dictionary. It is an ongoing project, implemented at the Department of Slavic Philology of Lomonosov Moscow State University in collaboration with the Department of Serbian Language at the University of Belgrade. For now, entries for words starting with the letters A, B, V, G, and D are already complete (approximately 10,000 entries in total), and those starting with the letters Đ, E, Ž, and Z are in progress. The estimated size of the complete dictionary is 80,000 entries.

The New Serbian-Russian dictionary is being created based on the most recent edition of the explanatory dictionary of the Serbian Academy of Sciences and Arts (RSJ, 2018). In compiling the dictionary and clarifying the meanings of words, text corpora, such as SrWac (2014) or PDRS (2022) are actively used. The dictionary is being developed in parallel in the traditional text format and in an electronic version on the OnLex platform (Makarov, 2024).

The article is organized as follows. Section 2 will present the principles of dictionary formation. Section 3 will be dedicated to the problem of lexical equivalence and the form of representing polysemy. Section 4 will introduce the digital version of the dictionary. Finally, Section 5 discusses the prospects for the development of the project.

2. Forming the Word List

We use the vocabulary, overall entry structure, and the meaning representation of the most recent Matica Srpska explanatory dictionary of Serbian (RSJ, 2018), comprising about 80,000 lexemes, as a starting point for our project. However, every part of a dictionary entry taken from RSJ (2018) is critically evaluated and the word list is recompiled taking into account the audience for which the new bilingual dictionary is intended.

The dictionary is devised primarily for Russian-speaking users. Its main purposes are to help the users read classic Serbian literature and contemporary texts of various genres, as well as to provide Russian learners of Serbian with the information necessary for the appropriate learning and everyday use of Serbian lexemes. These goals determine the principles of dictionary compilation, which we formulate as follows:

1. Lexemes that, according to the lexicographer's assessment, belong to the common lexical stock, are used in different functional styles, and are sufficiently frequent, are included in the Serbian-Russian dictionary as a matter of course.
2. Lexemes that are found in culturally significant literary works, both folk and authored, written in Serbian, are also included in the bilingual dictionary, even if their frequency in modern language is not high, e.g., *gradonosac* 'a cloud that carries hail'.
3. Words that belong to the terminological apparatus of a particular subject area are included in the dictionary if they are known to the modern educated speaker of the language; for more details on the principles of including terminological vocabulary, see Korolkova (2023).
4. International names of plants and animals are added to the dictionary if they meet two conditions:
 - a) the word has a different form in Russian, e.g., Serbian *ginseng* 'ginseng (Latin *Panax ginseng*)' corresponds to Russian *žen'shen'*, which looks and sounds differently.
 - b) a plant or an animal denoted by the term is important to either Serbian or Russian culture and everyday life, i.e., the term frequently appears in Serbian or Russian discourse.¹

¹ The same criterion applies to frequently used proper names which sound different in Serbian and Russian, cf. Serbian *Beč* 'Vienna,' corresponding to Russian *Vena*.

5. When selecting certain words, a word-formation criterion is used: a word is included in the dictionary if it serves as a productive base for various kinds of derivatives, even if the word itself is very rarely used in modern language (compare *Gestapo* ‘Gestapo’ — *gestapovski* ‘harshly’, *gonetati* ‘to guess’ — *zagonetka* ‘riddle’).
6. When including proper names in the dictionary, their semantic potential is also considered: if a proper name forms new meanings and can be used as a common noun, this is considered sufficient reason for including it in the dictionary (e.g., *Golgota* ‘Calvary’ and *golgota* ‘anguish, agony’).
7. In addition, the RSJ dictionary is supplemented with new words, primarily recent borrowings from English, e.g., *gejmer* ‘gamer’ or *globalizam* ‘globalism’. Dictionaries of neologisms, foreign words, and Anglicisms, such as Prčić et al. (2021) or Otašević (2008), are used as sources of such vocabulary, and the extent to which they are adopted, which is important for making the decision (not) to include a lexeme in the dictionary, is determined by their frequency.

The frequency of a word in modern Serbian is assessed using corpora of texts written in modern Serbian, such as PDRS (2022), SrWac (2014), SrpKor (2013) and some others. These corpora represent collections of modern internet texts collected from .rs domain sites over various periods between 2002 and 2022. All the corpora include a lot of data: the volume of the largest of the corpora, PDRS, is about 715 million tokens. We believe that such collections of texts provide a reliable representation of the degree of a word’s usage in the modern language; however, they certainly do not allow us to define the significance of the word for the general cultural and, especially, literary fund, as they do not take into account works written before the beginning of the 21st century. This lack of objective data is compensated for by the linguistic competence of the dictionary compiler since the word list is formed by a lexicographer who is a native speaker of Serbian.

3. Lexical Equivalence and the Problem of Representing Polysemy

When working with closely related languages, such as Serbian and Russian, the problem of lexical equivalence becomes particularly acute. Many words have a common origin and a very similar set of meanings. For example, the Serbian noun *bogatstvo* corresponds to the Russian cognate *bogatstvo*, and both words denote owning much property, as well as spiritual wealth (cf. Serbian *duhovno bogatstvo* vs. Russian *duxovnoe bogatstvo* ‘spiritual wealth’) and great variety (cf. Serbian *bogatstvo boja* vs. Russian *bogatstvo krasok* ‘a riot of color’). However, in most cases, careful analysis of collocations allows us to identify discrepancies in the use of tentative equivalents. For instance, the Serbian collocation *čitavo bogatstvo* ‘absolute fortune’ corresponds to the Russian expression *celoe sostojanie*, and not *’celoe bogatstvo*, as one might expect.

Large text corpora, including parallel ones, such as Serbian-Russian and Russian-Serbian fragments of the RNC (2024) or InterCorp (2023), provide us with the

opportunity to track subtle differences in superficially similar collocations and illustrate them with the most indicative, i.e., the most frequent and stable (idiomatic) examples. In addition, the possibility of publishing an electronic, not just a paper version of the dictionary allows us not to save space at the expense of illustrative material. New opportunities also change the general principles of developing a dictionary entry for a bilingual dictionary: a large amount of illustrative material allows for a focus on more precise explication of meaning, rather than on the selection of more precise translation equivalents, cf. Gudkov (1974).

In developing the New Serbian–Russian Dictionary, we are aiming for the comprehensive description of the semantics of the Serbian word. This corresponds to our general orientation towards Russian-speaking users, for whom Serbian is not a native language: it is precisely the representation of polysemy and possible contexts of word usage that facilitates subsequent correct use of Serbian lexemes and their adequate translation.

If a Serbian polysemous word has a Russian translation corresponding to it in the majority of senses (e.g., Serbian *gluv*, Russian *gluxoj* ‘deaf’), we do not just mention it in the entry. Instead, we try to present the entire structure of polysemy, i.e., to list all meanings of the Serbian word, illustrate each meaning with typical examples, and also include the synonyms of the translation specifically in this kind of usage. For instance, (1) illustrates the adjective *gluv*, which has several senses. In each sense, it modifies different nouns and has distinct translations. This way of representing polysemy allows us, among other things, to highlight the difference between the Serbian word and its primary translational equivalent (in this case, the cognate Russian *gluxoj* ‘deaf,’ which is given as a translation for all the meanings except 2d).

- (1) *A part of the entry for the Serbian gluv (the orthography of the dictionary is preserved; Serbian expressions are given in bold):*

1a. глухой, слабослышащий ‘deaf, hard of hearing’; ~ **на десно уво/ухо** глухой на правое ухо ‘deaf in the right ear’

1b. fig. глухой к чему-то, равнодушный ‘deaf, immune’; ~ **према нечијим жалбама, захтевима** глухой к чјим-то жалобам, требованиям ‘deaf to someone’s complaints or demands’

2. fig. a. глухой, тихий, без звуков; полный (о тишине) ‘quiet, dead, total (silence)’; ~**а ноћ** глухая ночь ‘dead of night’, ~**а тишина** глухая тишина ‘hollow silence’

2b. глухой, далекий, заброшенный, захолустный ‘remote’; ~**о село** глухая деревня ‘remote village’

2c. глухой, тихий, неживленный ‘deserted, quiet’; ~**а улица** глухая улица ‘quiet street’

2d. тихий, нешумный, без песен, веселья (о празднике) ‘quiet, noiseless, small’; ~ **свадба** тихая свадьба ‘quiet wedding’

2e. глухой, тихий, приглушенный, неясный ‘quiet, muffled, dull, hollow’; ~ **шум** глухой шум ‘dull sound, muffled noise’, ~ **глас** глухой голос ‘hollow voice’, ~и **корац**и глухие шаги ‘hollow footsteps’

The issue of explicating meaning (primarily of polysemous words but not only them) in creating a bilingual dictionary for closely related languages is the most crucial, and in our view, it has not yet received a definitive solution. As new data accumulate, we refine and revise our principles for organizing dictionary entries and representing polysemy, and, just as importantly, we notice various patterns in the structure of the lexical systems of closely related languages that were not visible before.

4. OnLex Platform

Aside from publishing a hardcover edition of the dictionary, we are also developing its online version, powered by the OnLex platform (Makarov, 2024). The main motivations include being able to make constant amendments after the physical copy is ready and increasing the availability of the dictionary. Not only the OnLex platform is a publishing solution, allowing one to search within different parts of the entry and therefore making the dictionary useful for Serbian learners of Russian as well, but it is also advanced dictionary writing software, providing a lexicographer with a means of editing every component of the entry via an intuitive user interface. Finally, the online dictionary can render supplementary media files (e.g., we are planning to add audio files illustrating the pronunciation of lexeme paradigms since Serbian has variable lexical stress) and include more examples.

The screenshot shows the OnLex platform interface for the entry 'азбука'. At the top, there are tabs for 'Формы и значения', 'Примеры', 'История', 'женский род', and 'Создать задачу'. The main content area is divided into sections: 'Значение' (Meaning), 'Устойчивые сочетания и идиомы' (Stable combinations and idioms), and 'Оригинальный текст статьи' (Original article text). The 'Значение' section lists the word as a noun (сущ.) and provides examples in Russian, Greek, and Braille. The 'Устойчивые сочетания и идиомы' section lists 'Морзеова' and 'Брајева' with their meanings. The 'Оригинальный текст статьи' section contains the original text of the article.

Fig. 1: Screenshot of the entry *azbuka* ‘alphabet’ as rendered by the OnLex platform. In the upper left corner word forms are given (the main form, labelled ‘зс,’ and a dative form, labelled ‘дат.’); note that in the original text of the entry (‘Оригинальный текст статьи’), i.e., in the print edition, the dative form is shortened. To the right of the word forms, there are senses of the lexeme (‘Значение’), illustrated with two examples and their translations, and idioms / fixed expressions (‘Устойчивые сочетания и идиомы’) along with their meanings. Every abbreviation, be it a form type (зс = headword, дат. = dative), grammatical information (ж = feminine, сущ. = noun) or a usage label, is given a pop-up tooltip, i.e., the full form appears on the screen if the user positions a cursor over the abbreviation.

As shown in Figure 1, every part of the entry is represented separately, i.e., the OnLex platform treats it as a distinct entity and stores it in a separate cell in the database. Such an approach allows us not only to render each entry component with a distinct set of typographical conventions but also to automatically generate a layout required for the paper edition. On top of that, the database structure treating every part of the entry as a separate entity (rather than representing the whole entry as a string of characters with no markup) allows for flexible search queries. For example, it is possible to search not only within headwords but also within idioms, examples, word forms (e.g., dative forms), etc.

The screenshot shows the OnLex entry editing interface. It consists of two main sections: 'Значения заглавной лексики' (Main Lexical Values) and 'Устойчивые сочетания и идиомы' (Stable Collocations and Idioms). The first section has a table with columns: Ранг (Rank), Пометы (Labels), and Значение (Meaning). The second section has a table with columns: № (Number), Пометы (Labels), Формы (Forms), Ранг (Rank), Пометы (Labels), and Значение (Meaning). Arrows point to specific elements: 'Rank' points to the 'Ранг' column in the first table; 'Drop-down menu for part of speech' points to the 'Пометы' column in the first table; 'Grammatical labels' points to the 'Пометы' column in the second table; and 'Drop-down menu for other labels' points to the 'Пометы' column in the first table.

Fig. 2: Screenshot of the entry editing interface used by the OnLex platform with the main elements labelled. Ranks are used to order entry elements. Each label is chosen from predefined lists, which minimizes the chance of misprints.

The OnLex entry editing interface is partly shown in Figure 2. Note that no programming or technical skills are required to fill in the fields of this interface – editors either have to type unformatted text or select the required options from drop-down menus with predefined labels. All the changes are implemented and available for users right after the editor clicks on the save button and the system keeps the record of the versions of the entry. This interface is also used for writing new entries, which can be added to the online version immediately or just saved for internal use.

Among other features of the OnLex platform, there is a task tracker tailored specifically to the needs of the lexicographers working on the dictionary. Tasks can be created for specific entries and/or assigned to specific editors. These options make it possible for a logged-on editor to see that a particular entry has been mentioned in a task and when the task is assigned to a particular editor, this editor is notified. For each entry, there is a discussion board and a timeline showing who and when worked on the task.

5. Conclusion and Future Prospects

In this article, we have introduced the New Serbian-Russian Dictionary, the creation of which is associated with solving a number of practical and theoretical problems. The main theoretical problem facing the compilers of such a dictionary is the issue of the presence or absence of equivalence between translation units. This question is related to the means of representing polysemy: is it sufficient to bring the closest translation equivalent if it has a similar structure of polysemy, or should the set of meanings of a polysemous word be presented in detail in any case, supplying it with numerous illustrations? In the current version of the dictionary, the second option is implemented, but as we accumulate material, we constantly modify and refine the principles we have already developed.

Practical tasks are associated with the choice of methods for analyzing and presenting material. In analyzing the meanings of words and searching for the most illustrative examples, we use a whole range of available corpora; however, each of them reflects only the contemporary state of the language and does not include classical literature. Using the online platform OnLex contributes to more flexible work on the dictionary. It facilitates the task of adding new data and the process of changing the existing dictionary entries, as well as makes it possible to attach an unlimited number of illustrations and full paradigms. It is worth pointing out that each entry can be supplemented with audio recordings of each form constituting the lexeme's paradigm, which is important for languages with variable (i.e., not fixed) lexical stress.

Use of the methods of online lexicography opens up great prospects for further work: it is possible to reflect ongoing lexical changes in both languages via regular and systematic updates of the bilingual dictionary, to add translations of the same dictionary entries into other languages (for example, other Slavic languages), to conduct various kinds of quantitative research, and produce different visualizations (e.g., graphs illustrating relations between translation equivalents).

References

- Gudkov, V. P. (1974). Dvujazychnye slovari rusko-serbskokhorvatskie, serbskokhorvatsko-russkie i perspektiva ikh sovershenstvovanija [Russian-Serbo-Croatian, Serbo-Croatian-Russian bilingual dictionaries and the prospect of their improvement]. *Sovetskoe slavianovedenie*, 6, 65–73.
- InterCorp 2023: *The parallel corpus InterCorp. Release 16*. <https://intercorp.korpus.cz/?lang=en>, (last access: 1 June 2024)
- Korolkova, M. D. (2023). Special'naja leksika v dvujazychnom slovaru [Special vocabulary in a bilingual dictionary]. *Slavistika*, 27(2), 190–201.
- Makarov, Y. (2024). Principles and methods of digital lexicography. *Izvestiia Rossiiskoi akademii nauk. Seriya literatury i iazyka*, 83(4), in print.
- Otašević, Đ. (2008). *Rečnik novih reči* [Dictionary of new words]. Alma.

PDRS 2022: *Serbian Web Corpus PDRS 1.0*. <https://javnidiskurs.rs/veb-korpus-serbian-web-pdrs/>. (last access: 1 June 2024)

Prčić, T., Dražić, J., Milić, M. et al. (2021). *Srpski rečnik novijih anglicizama* [Serbian Dictionary of new anglicisms]. Filozofski fakultet.

RNC 2024: *Russian National Corpus*. Retrieved June 1, 2024, from <https://ruscorpora.ru/>

RSJ 2018: *Rečnik srpkoga jezika*. Matica Srpska.

SrWac 2014: *Serbian Web Corpus v1.2*. Retrieved June 1, 2024, from <https://ruscorpora.ru/>

SrpKor 2013: *Korpus savremenog srpskog jezika*. Retrieved June 1, 2024, from <https://ruscorpora.ru/>

Tolstoy, I. I. (1970). *Serbskokhorvatsko-russkii slovar'* [Serbo-Croat–Russian dictionary]. Sovetskaja entsiklopedija Publ.

Acknowledgements

We deeply thank our anonymous reviewers for their valuable questions and useful comments. The study is funded by the Russian Science Foundation, project No. 23-28-00126, <https://rscf.ru/project/23-28-00126/>.

Contact information

Rajna Dragičević

University of Belgrade, Belgrade, Serbia
rajna.dragicevic@fil.bg.ac.rs

Yury Makarov

University of Cambridge, Cambridge; Vinogradov Russian Language Institute, RAS; Institute of Linguistics, RAS
im562@cam.ac.uk

Daria Ryzhova

HSE University, Moscow, Russia
daria.ryzhova@mail.ru

Yulia Shapich

Lomonosov Moscow State University, Moscow, Russia
julija.sapic@gmail.com

Ekaterina Yakushkina

Lomonosov Moscow State University, Moscow, Russia
jkatia@yandex.ru

Mariusz Piotr Kamiński

THE CONTRIBUTION OF BILINGUALIZED ENTRIES AND VOCABULARY KNOWLEDGE TO THE LEARNER'S SUCCESS IN SENTENCE COMPLETION

The Case of *jump* Verbs

Abstract This study examines the role of the BLD and the user's receptive vocabulary knowledge in making appropriate lexical decisions in a sentence completion task. Fifty-two advanced and low-proficiency learners were recruited for the study. They were asked to read a set of gapped sentences and fill in the blanks with semantically similar verbs, using a randomly assigned dictionary (either BLD or MD depending). A logistic mixed-effects regression model was developed in which a participant's choice of a verb was treated as a success or failure in the task. The model showed that the successful completion of the task significantly depended on the learner's vocabulary knowledge rather than on the type of dictionary. Likewise, another regression model (two-way ANOVA) revealed that the time taken to complete the task was significantly affected by learners' vocabulary knowledge but not by dictionary type.

Keywords bilingualized dictionary; equivalent; vocabulary; mixed-effects

1. Introduction

The learner's bilingualized dictionary (BLD) offers the best features of both mono- and bilingual dictionaries: an L2 definition and an L1 equivalent of the headword. The rationale behind this type of dictionary is to assist learners in understanding the English definition by providing them with "limited help in the mother tongue" (Reif, 1987, p. 153). An example of this approach is the entry for *crease* from an English-Polish dictionary (Phillips, 2002):

crease ... noun [c] **1** an untidy line on paper, material, a piece of clothing, etc. that is caused by not treating it carefully: *Your shirt needs ironing, it's full of creases* ... **zmarszczka, zagięcie**

In this entry, the meaning of *crease* is rendered by the English definition and example sentence, and a few equivalents in the learner's native language (Polish). An equivalent provided in such a dictionary can save the user a great deal of time, especially when it carries the same semantic overtones as the word defined. The role of the equivalent is to reassure the learner about the meaning conveyed by the definition (Abecassis, 2007). The risk of misunderstanding the meaning appears to be high in those definitions that are written with a controlled vocabulary, since they are claimed to be inefficient in differentiating finer shades of meaning (Stein, 1979). Grasping meaning differences between semantically similar words is crucial for learners striving to write accurate sentences.

However, contrary to the above assumptions, previous empirical studies (Lew, 2004; Kamiński, 2023) suggest that the effect of the BLD in word-in-context selection tasks is not as substantial as expected in comparison with the monolingual dictionary (MD). The present study examines whether using BLDs can make sentence completion task easier for a learner compared to MDs, and whether success in this task depends on the learner's vocabulary knowledge. Unlike Kamiński's earlier study (2023), the present research focuses on a different set of words and uses a different statistical method (mixed-effects model).

2. Research Questions

This study tested the effectiveness of the BLD in a gap-fill task requiring participants to make decisions on the choice of verbs to complete sentences. Each gap could be filled only by one of the near-synonyms for *jump*. The focus was on two research questions:

Do dictionary type (BLD or MD) and learners' vocabulary knowledge influence:

1. the appropriateness of lexical choices in this task?
2. the time taken to complete the task?

3. Method

The research questions were addressed by building two regression models, in which dictionary type (BLD and MD) and learners' vocabulary level (advanced and low) served as predictors.¹ The first model, a mixed-effects model, measured the appropriateness of lexical choice as the outcome variable (appropriate or inappropriate), while the second model measured the time taken to complete the task. The data were collected through a user-oriented study, in which each student gave repeated responses to a different sentence context.

3.1 Participants

Fifty-two undergraduate students of English Studies at the University of Applied Sciences in Nysa (Poland) participated in this research. They were all native speakers of Polish, with a minimum of 11 years of formal English instruction in schools within the Polish school system. Their level of English was estimated using a receptive vocabulary size test (Nation & Webb 2011). The test showed that participants' knowledge of vocabulary ranged between 4,500 and 14,400 word families, with the median being 9,750. Those learners whose vocabulary size was below 9,000 word families were classified as low-proficiency learners, while the others as advanced learners. Participants were randomly divided into two groups: one with a BLD, and the other with an MD at their disposal.

¹ The data preparation and statistical analyses were performed in R (R Core Team 2023).

3.2 Materials and Procedure

The task included 7 sentences from a proficiency-level coursebook (Stephens, 2002, p. 21), each with a gap to be filled with one of the target verbs: *creep, crouch, leap, lurk, pounce, scramble, stalk*. The verbs were potentially unknown to participants since they are rare in the language, and not included in their course curricula. An additional sentence, where none of the words fit, was added to minimize the effect of random guessing. Participants could earn a total of eight points, leaving one gap blank. The task featured a set of dictionary entries for the verbs, available in two versions: MD or BLD, depending on the student group.

The two versions shared identical monolingual content drawn from the *Longman Dictionary of Contemporary English, Writing Assistant Edition CD-ROM* (LDOCE4; Summers, 2005), except for the learners' L1 equivalents (drawn from *The New Kosciuszko Foundation Dictionary*; NKFD; Fisiak, 2003) exclusive to the BLD. An example of the BLD entries, along with the task sentence, is shown in Figure 1. The MD version of this entry had no right-hand panel. The entries were reduced to relevant senses only. The majority of the equivalents were one-word lexical units. Students completed the task, using either the MD or BLD entries depending on the assigned group. The task was administered online, using an HTML and PHP interface, with the task sentences and entries displayed on the same webpage. The experiment was conducted in a university computer room during normal class time. Learners' responses and completion times were collected in a MySQL database. Following the main task, participants were asked to take the vocabulary size test, which was also available on the web (VocabularySize.com, 2024).

The tiger a small antelope through the long grass.

stalk	[transitive] to follow a person or animal quietly in order to catch and attack or kill them [shadow]: <i>Polar bears stalk seals that are resting on the ice.</i> <i>We know the rapist stalks his victims at night.</i>	śledzić, tropić; podchodzić (zwierzę, ofiarę).

Fig. 1: One of the task sentences and the corresponding BLD entry for *stalk*. The entry combines the monolingual English part (from LDOCE4) and the L1 equivalents (NKFD)

4. Results

4.1 Appropriateness of Word Choice

To address the first research question, a logistic mixed-effects model was developed, where each participant's lexical decision resulted in either success or failure (outcome variable). The model formula notation was as follows:

`glmer(Appropriateness ~ Dictionary + Vocabulary + ( 1|ID) + (1|Item), data = data, family = "binomial", control = glmerControl(optimizer = 'Nelder_Mead'))`

In this model, Appropriateness stands for the outcome variable, Dictionary and Vocabulary are the predictors, ID is a participant, and Item stands for a verb.

Table 1 outlines the learners' mean scores. Each participant could score a maximum of 8 points for accurate responses. The scores indicate that the BLD users made on average more appropriate lexical decisions than the MD users. Taking into consideration their vocabulary knowledge (Table 2), one can see that this variable seems to be an important factor determining learners' success in the task. Advanced learners performed better than low-proficiency learners, irrespective of the dictionary used.

Table 1: Descriptive statistics for the appropriateness of word choice

Dictionary	N	Mean	SD
MD	33	5.33	2.07
BLD	19	6.05	2.09

Table 2: Appropriateness of word choice by learners' vocabulary level and dictionary type

Dictionary	Vocabulary	N	Mean	SD
MD	advanced	22	6.0	1.57
	low	11	4.0	2.36
BLD	advanced	13	7.15	1.06
	low	6	3.66	1.75

The results of the logistic model are shown in Table 3. Of the two predictors, only the learners' vocabulary level ($\beta = -2.16$, $p < 0.0001$) had a significant effect on whether they chose verbs appropriately or not. In other words, for a learner with a limited vocabulary size, as compared to an advanced learner, the odds of being correct in this task decrease by a factor of 8.67 ($e^{2.16}$).

Table 3: Model estimates and p -values

	Estimate β	SE	z-value	p-value
Intercept	1.78	0.52	3.44	< 0.001
Dictionary[BLD]	0.79	0.47	1.67	0.09
Vocabulary[low]	-2.16	0.48	-4.53	< 0.0001

The model assumed a lack of interaction between Dictionary and Vocabulary. A model that allowed the interaction was also fitted but the interaction was not significant. A likelihood ratio test comparing the two models showed that neither of them appeared to be better than the other ($\chi^2(1) = 3.1$, $p = 0.08$). Another likelihood ratio test showed that the fitted model is significantly better than a simple logistic model without random effects (AIC = 410 vs. 468). It accurately predicts the appropriateness of word usage 85% of the time in the data modeled. The predicted probabilities of appropriate word choice are illustrated in Figure 2. As can be seen, only advanced learners are

predicted to provide successful responses in this task, while the probability of success for low-proficiency learners is less than chance. This means that dictionary type is a rather poor predictor of the task success.

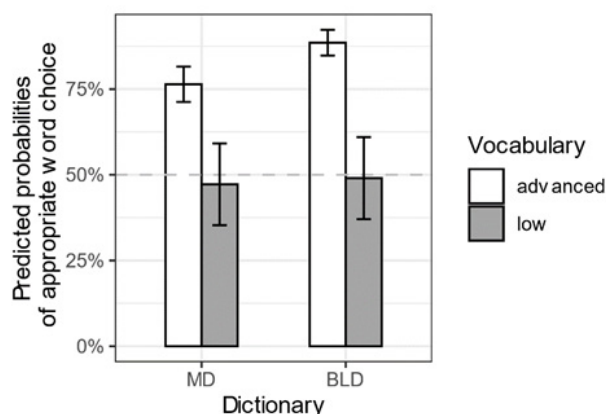


Fig. 2: Predicted probabilities of appropriate word choice by dictionary type and vocabulary level

4.2 Time Taken to Complete the Task

With respect to the second research question, a two-way ANOVA was conducted to estimate task completion time based on the same predictor variables: Dictionary and Vocabulary. The model allowed for an interaction between the predictors, as suggested by the distribution of times in Figure 3. The figure shows that only advanced learners took advantage of the BLD, since they completed the task faster than less competent learners, and, as indicated in Table 2, they achieved higher scores in word selection (7.15 vs. 6.00). However, the effect of dictionary on time was not significant ($p > 0.05$), according to the model (see Table 4). Table 4 indicates that time is significantly influenced by learner's vocabulary level ($p < 0.05$, $\omega^2 = 0.08$). Advanced learners need less time to fill in sentence gaps than intermediate learners.²

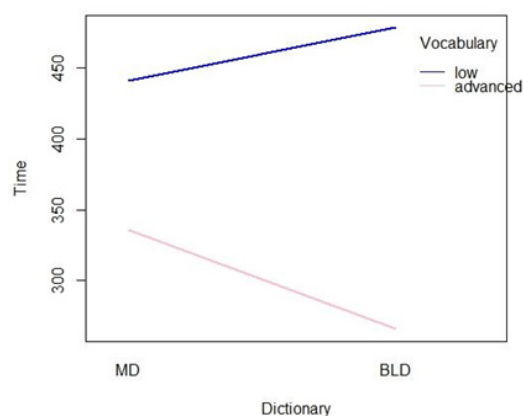


Fig. 3: Entry consultation time (in seconds) by dictionary type and learners' vocabulary level

² The design of this study did not allow us to separate the actual entry consultation time from the task completion time, as both activities are typically done concurrently. It follows that one cannot rule out the possibility that some learners, particularly the very advanced ones, filled the sentence gaps without consulting the entries. Using a more sophisticated approach, such as eye-tracking, would provide insight into whether this was the case.

Table 4: The ANOVA output estimates and *p*-values

	<i>Df</i>	Sum Sq	Mean Sq	<i>F</i> value	<i>p</i> -value
Vocabulary	1	156661	156661	5.53	0.02
Dictionary	1	820	820	0.03	0.87
Vocabulary:Dictionary	1	2033	2033	0.07	0.79
Residuals	48	1360573	28345		

4.3 Conclusions

This study explored the contribution of dictionary type and learners' vocabulary knowledge to making appropriate lexical choices in the sentence completion task. The following are the answers to the research questions:

1. Learners' vocabulary knowledge significantly contributes to making appropriate decisions on the choice of verbs for sentence gaps. Advanced learners performed significantly better than low-proficiency learners. The task success was not significantly affected by whether participants used BLD or MD.
2. The time taken to complete sentences was significantly affected by learners' vocabulary knowledge, with advanced learners choosing the verbs faster than the less competent students. No significant effect of dictionary type on the task completion time could be shown.

From the findings, it transpires that the knowledge of English vocabulary is more crucial than the choice between a BLD or an MD in tasks requiring the ability to distinguish between semantically similar verbs. The findings not only underline the importance of learners' profiles, in this case their proficiency level, but also cast light on the rather limited role of BLDs in this particular task. In evaluating this role, one cannot evaluate the usefulness of the dictionary in separation from such factors as the learner's profile and the type of task at hand. The experiment was conducted on a modest selection of verbs and with one particular pair of languages, so further research is needed to see whether the findings can be generalized to a wider population.

Finding (1) supports the conclusion reached in the earlier study by Kamiński (2023) that neither the BLD nor the MD shows a significant advantage over the other in the word-in-context selection task. However, in contrast to the previous study, finding (2) revealed that the learner's vocabulary had a significant impact on task completion time.

References

Abecassis, M. (2007). Is Lexicography Making Progress? On Dictionary Use and Language Learners' Needs. *Lexikos*, 17, 247–258. <https://doi.org/10.5788/17-0-555>

Fisiak, J. (Ed). (2003). *The New Kosciuszko Foundation Dictionary English-Polish and Polish-English*. The Kosciuszko Foundation.

Kamiński, M. (2023). How Useful Are Bilingualized Dictionaries in Discriminating Between Near-Synonyms? *International Journal of Lexicography*, 36(4), 486–509. <https://doi.org/10.1093/ijl/ecad019>

Lew, R. (2004). *Which dictionary for whom? Receptive use of bilingual and semi-bilingual dictionaries by Polish learners of English*. Motivex.

Nation, I. S. P., & Webb, S. (2011). *Researching and Analyzing Vocabulary*. Heinle.

Phillips, J. (Ed.). (2002). *Oxford Wordpower. Słownik Angielsko-Polski z Indeksami Polsko-Angielskim*. Oxford University Press.

R Core Team. (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. <https://www.R-project.org>.

Reif, J. A. (1987). The Development of a Dictionary Concept: An English Learner's Dictionary and an Exotic Alphabet. In A. P. Cowie and European Association for Lexicography (Eds.), *The Dictionary and the Language Learner. Papers from the EURALEX Seminar at the University of Leeds, 1–3 April 1985* (pp. 146–158). Max Niemeyer Verlag.

Stein, G. (1979). The Best of British and American Lexicography. *Dictionaries* 1, 1–23.

Stephens, M. (2002). *New Proficiency Reading*. Pearson Education Limited.

Summers, D. (Ed.). (2005). *Longman Dictionary of Contemporary English, Writing Assistant Edition CD-ROM* (4th edition). Pearson Education Limited.

VocabularySize.com (2024). *VS Test Your Word Knowledge*. Retrieved March 2, 2024, from <https://my.vocabularysize.com/>

Contact information

Mariusz Piotr Kamiński

Department of Foreign Languages, University of Applied Sciences in Nysa, Poland
mariusz.kaminski@pans.nysa.pl

Evelīna Ķiršākmene

FALSE FRIENDS IN GENERAL BILINGUAL DICTIONARIES (ENGLISH AND FRENCH INTO LATVIAN)

Abstract Over the last few decades, the influence of the English language onto the Latvian language has significantly increased. This is due to geopolitical events, rapidly growing media and technological advances where the global language is English. Consequently, the influx of English borrowings into Latvian has increased, raising the incidence of false friends. Notably, some long-established false friends either adopt another meaning or lose the original one aligning more closely with their English counterpart. French has also impacted Latvian, albeit through intermediary languages such as German and Russian. In this paper, examples of false friends in English-Latvian and French-Latvian bilingual dictionaries will be viewed and compared. Currently, there is one English-Latvian false friend dictionary published over three decades ago; the necessity for revisions that would account for recent linguistic developments need to be determined. Contrastive lexicographic analysis is applied.

Keywords false friends; borrowings; diachronic change; lexicographic treatment

1. Introduction

There are numerous definitions of false friends with the focus point of them being ‘words within two different languages which are graphically or phonetically very similar but have different meanings and can therefore be easily confused by foreign language learners’ (Gorbahn & Hausmann, 1989, p. 2882). This definition outlines the possible problems in language acquisition, but they can create problems in translation, interpreting and general communication as well. Often the terms *false friends* and *false cognates* are used as synonyms (Hartmann & James, 1998, p. 56), but some linguists strongly distinguish them and propose the use of *false cognates* and *deceptive cognates* only (Sabino, 2016). In the present paper, the term *false friend* will be used as a superordinate term.

2. Current Situation of False Friends in Latvian

Lexicographers mainly focus on semantic false friends: cognates that either have completely different meanings in each language (monosemantic) or are highly polysemantic in the source language, but have significantly fewer meanings in the other language (polysemantic), thus narrowing the semantic scope and becoming a false friend. Attention to chance false friends with no etymological links is not usually as high on agenda since it is believed they pose less issues than semantic false friends. Veisbergs also mentions pseudo false friends made by language users creating non-existent words in a foreign language imitating their native tongue words (Veisbergs, 1996, p. 629).

2.1 Borrowings, Linguistic History

The Latvian language has a high number of borrowings because of protracted and extensive direct language contacts with German, Finno-Ugric languages, Polish, Swedish, and Russian. Due to historic reasons, most borrowings come from Russian and German, which have also served as intermediary languages for loans from other languages, like English and French. Now, however, many German borrowings as well as Russian ones have been downgraded to lower register. The current geopolitical situation is also marking a change in the role of the Russian language in Latvia, it being taken out of education programs among other changes introduced.

Even though the possibility of false friends is higher among kindred languages, Latvian, which is a Baltic language, has numerous false friends from English and French. These combinations present interest in two aspects: firstly, French was often the language behind the loans, with meaning already changed in the intermediate language. A few examples of this could be the French words ‘agitation’, ‘aspirant’, ‘intelligence’ – all have a different meaning in Russian or German and, thus, also in Latvian. Secondly, some have undergone semantic change aligning the meanings with the intermediary languages and some have developed specifically Latvian nuances.

It is considered that during the Soviet times the Latvian language was more controlled in the media and literature, therefore false friend mistakes were less frequent (Baldunčiks, 2006). However, it is a myth that the real language used to be purer in the past. This paper focuses on false friends from English and French, but if we viewed Russian and German false friends, it is very likely that we would find a similar situation.

2.2 Emerging Diachronic Changes

Since English as the global communication language has had a huge impact on Latvian during the last decades, it frequently affects the semantics of Latvian words. Some false friend meanings are merging, and words are used in such a context that previously would have been considered wrong. Some semantic change mechanisms include specialization or narrowing and generalization or widening (Broz, 2008, p. 205). This is showcased with the word *dekāde* which was previously mainly used with one stable meaning in Latvian: ‘a period of ten days.’ Now, it is increasingly being used with its English counterpart meaning ‘ten years’, as it can be observed in the Balanced Corpus of Modern Latvian (LVK2022).

Another example would be *afēra* [*affair* in English, *affaire* in French], which is a highly polysemantic word both in English and French. It had only one meaning in Latvian, that of ‘a dishonest, speculative venture, a fraudulent transaction’ (MLVV). Now, it is also being used with the meaning ‘love affair’ – the same as in English. This shows expanding polysemy and semantic broadening. The English language influence is manifested as adoption of the English meaning parallel to the Latvian one. This leads us to the treatment of such semi transient false friends in bilingual dictionaries and on the most appropriate ways to reflect these ongoing semantic changes.

3. False Friends in Latvian Bilingual Dictionaries

False friends constitute a small share of the overall dictionary entry numbers, but their treatment deserves close attention. The question of equivalent validity is of importance. It has been stated that since the studies of false friends have appeared relatively recently, mistakes of their translation in bilingual dictionaries are possible (Veldi, 2006). It depends on the false friend type, but the dictionary should not delude the user into thinking that the similar counterparts in both languages are complete equivalents. This leads to several possible suggestions on how to best inform the dictionary user of false friends.

Often, specific information on false friends is not incorporated in bilingual dictionaries at all, them not being distinguished apart in any way. This leads to false friends “disappearing” in the large mass of words with the only treatment strategy being correct equivalent, example provided if possible, and abbreviations of labels to indicate the field of use. Some imply the use of a specific symbol of false friends in dictionaries (Gouws, Prinsloo & de Schryver, 2004, p. 799). Even though every additional symbol or mark would increase the dictionary size, perhaps this would lead to saving some space at the end, as it would erase the need to use additional examples to explain the semantic differences. However, it is not clear whether only informing the user about a false friend would be sufficient; most likely, some indications would be needed to explain the difference. Even more so, the question of marking would arise – should all false friends be marked or the most frequent or popular ones? Would the mark not complicate perception?

Space limitations in paper dictionaries are considered the main obstacle when dealing with false friends since they are so complex in nature (Granger & Swallow, 1988). Digital dictionaries do not possess this obstacle and could theoretically have lengthy dictionary entries. But there are counterarguments to this as well: ‘it would make looking up a fairly simple entry a more complicated task simply because of the amount of information, and it would obliterate the notion of lexical set’ (Corréard, 2002, p. 466). This leads to the conclusion that treatment of false friends in general dictionaries is complicated; there are advantages and disadvantages to every solution or strategy. Veldi states that “comprehensive treatment of false friends is possible only in specialized dictionaries where various meanings can be adequately compared and contrasted” (Veldi, 2006, p. 179). Perhaps a starting point would be providing correct equivalents of false friends in bilingual dictionaries, i.e., not putting an equivalence mark between the two and creating the illusion of both words meaning the same. Labels can be used to underline different domains of word use. It is not known whether it would be an efficient practice to include a specific statement at the end of the lemma informing of the specific false friend and explaining the differences. An illustration of this example could look like this: “translation: tulkojums [translation], not translācija [broadcast]!”.

In this paper, false friend examples in English-Latvian and French-Latvian bilingual dictionaries are analysed. They have been randomly chosen from the false friend corpus currently being collected by the author consisting in total of 194 false friends,

32 of which are chance false friends, 65 monosemantic and 97 polysemantic false friends. 115 examples or 59 percent of the total are English-Latvian false friends, and 79 examples or almost 41 percent are French-Latvian examples.

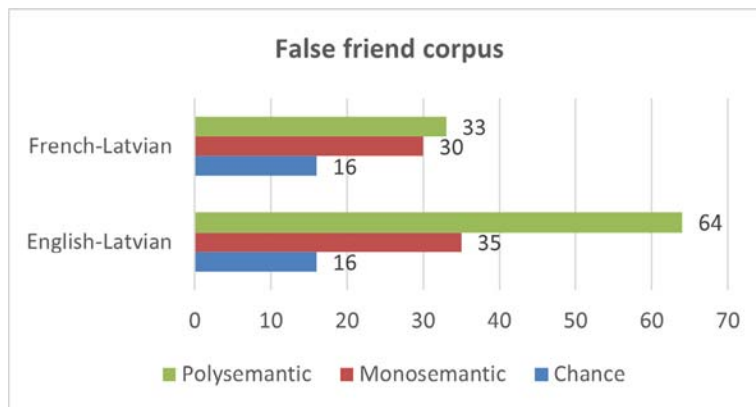


Fig. 1: Information on the false friend corpus collected by the author

The samples provide an insight in the treatment of false friends in Latvian bilingual dictionaries analysed during this paper. Four English-Latvian and three French-Latvian dictionaries were reviewed, which are the most recent dictionaries available. The digital dictionary *Tilde*, which is one of the few Latvian digital dictionaries available online, was used in both language combinations. *Tilde* is based on paper dictionaries.

The definitions used to differentiate false friend meanings in each language have been chosen from the *Mūsdienu latviešu valodas vārdnīca* [Contemporary Latvian language Dictionary] (hereinafter – MLVV), *Oxford Learner's Dictionaries* (OLD) and *Larousse French dictionary* (Larousse). Two English-Latvian false friend examples will be analysed first, then moving on to two examples from French into Latvian.

1. In English, *anecdote* can be a joke and ‘a personal account of an event’ (OLD). *Anekdotē* in Latvian signifies ‘a short story about a comic event with an unexpected and witty ending’ (MLVV). It is a polysemantic false friend with two respective meanings in English and one in Latvian. Below is the information available at the English-Latvian dictionaries. The translation is provided in brackets next to the entry.

Table 1: Comparison of English-Latvian dictionary entries of “anecdote”

ENG 1	ENG 2	ENG 3	ENG 4
anekdotē [a joke]	anekdotē [a joke]	1.anekdotē [a joke]; 2.īss stāsts, epizode, asprātīgs stāsts [a short, witty story, episode] 3.privātās dzīves sīkumi par slavenību (parasti npublicēti) [details of celebrity private life (usually unpublished)]	anekdotē [a joke]

Only one meaning is provided without mentioning semantic nuances in three out of four dictionaries. Therefore, the language user might automatically presume that words in both languages carry identical semantic weight, which is not true. The situation is similar with other false friend examples like *exploitation* [ekspluatācija] and *utilisation* [utilizācija], which all have significant meaning differences that are not indicated in dictionaries.

2. In English, *gymnasium* designates a sports hall. Ģimnāzija in Latvian is a loan from German, referencing to a type of a secondary school starting from the 7th grade until the 12th with higher education standards and requirements. In some countries, the meaning of this word coincides with the Latvian meaning, but the first equivalent of this word from English into Latvian should be ‘a gym’.

Table 2: Comparison of English-Latvian dictionary entries of “gymnasium”

ENG 1	ENG 2	ENG 3	ENG 4
1.vingrošanas zāle [gym]; 2.ģimnāzija [secondary school]	1.ģimnāzija [secondary school]; 2.vingrošanas zāle [gym]	1.vingrošanas zāle [gym]; 2.ģimnāzija [secondary school]	1.vingrošanas zāle [gym]; 2.ģimnāzija [secondary school]

As we can see, all dictionaries provide two meanings; three offer the correct first equivalent. The information and context provided is indeed very limited; it might not be clear to the language user which meaning applies in the specific case. This example also showcases a lack of subject labels or explanations of differing word meanings in different countries. The same with the false friend *partisan* [partizāns], which is being used in different contexts both in English and Latvian.

Now some examples in French-Latvian dictionaries:

3. A polysemantic false friend is the French word *collaboration*. Its semantic spectrum is broad with the first meaning being ‘participation, cooperation’ (Larousse). *Kolaborācija* in Latvian is used with negative connotations, referring to collaboration with the enemies of the state, referencing mainly historical events of the 20th century. Results suggest that the field of use is indicated for the agent noun (explaining that it is a political term), but without the explanation of the historical and political context to what it refers to.

Table 3: Comparison of French-Latvian dictionary entries of “collaboration”

FR 1	FR 2	FR 3
līdzdarbība [complicity]; collaborationiste <i>m. pol.</i> kolaboracionists [collaborationist]	1.piedalīšanās [participation]; sadarbība [cooperation]; 2. <i>hist.pol.</i> sadarbošanās [collaboration]	līdzdarbība [complicity]; collaborationiste <i>m. pol.</i> kolaboracionists [collaborationist]

All three sources provide additional labels to indicate a specific domain of use which proves to be a useful practice. The polysemantic false friend *audience* [audience] with numerous meanings in both French and Latvian was treated in dictionaries similarly. The same with *diète* [diēta] with two meanings in French but only one in Latvian: two distinct subject labels solved the ambiguity.

4. *Démonstration* in French has seven listed meanings in Larousse, starting from ‘scientific proof or argument’ and ‘manifestation’ to ‘a display of emotions’. *Demonstrācija* has two meanings in Latvian: a display of something and ‘the expression of public sentiment, political demands or protest in a mass march, rally’ (MLVV), making it a polysemantic false friend. Below are the results obtained from French-Latvian dictionaries.

Table 4: Comparison of French-Latvian dictionary entries of “démonstration”

FR 1	FR 2	FR 3
1. <i>mat.</i> pierādījums [proof]; 2. <i>izrādīšana</i> ; <i>parādīšana</i> [display]; 3. <i>daž. noz. [different domains]</i> demonstrēšana; demonstrācija [demonstration]	1. <i>izrādīšana</i> [display]; 2. <i>demonstrējums</i> [demo]; 3. <i>pierādījums</i> [proof]; 4. <i>spēka/drosmes izrādīšana</i> [manifestation of force/courage]	1. <i>mat.</i> pierādījums [proof]; 2. <i>izrādīšana</i> ; <i>parādīšana</i> [display]; 3. <i>daž. noz. [different domains]</i> demonstrēšana; demonstrācija [demonstration]

Though several meanings appear in the Latvian lemma, *demonstrācija* with its multiple domain label will look the most attractive Latvian counterpart to the user. Here again labels are used to provide additional information to the dictionary user. The provided equivalents could necessitate some examples in context to understand how these words really function. But that, of course, is a question of space – since only one of the six dictionaries analysed in this paper is digital, this is an important factor that must be taken into consideration.

4. False Friend Dictionaries in Latvian

False friend dictionaries constitute a specific lexicographical category. Their target audience would be language professionals like interpreters, translators, language teachers and others. Therefore, these types of dictionaries have specific requirements: accidental false friends should be excluded; specific attention needs to be paid to the selection of headwords; the offered equivalent should be correct with clear examples, reflecting the present-day usage of the word; difference explanations should not be long but to the point, and others (Gorbahn & Hausmann, 1989, p. 2883). The field of use, register level and possible idiomatic meanings of the false friends should be indicated separately.

Currently, there is one false friend dictionary in the English-Latvian language combination by Andrejs Veisbergs (1994). It consists of approximately 800 entries. Only semantic false friends are included, excluding chance and pseudo false

friends. The domain of use and explanation is provided in the brackets to facilitate comprehension. Both Latvian and English false friends and their correct equivalents are in columns next to each other which helps see semantic differences visually.

A study of French-Latvian false friends was carried out by Andrejs Bankavs (1989), offering analysis of false friends on a morphological, lexical, and syntactical level. Here the goal was not to compile as many false friends as possible, but to explain differences in French and Latvian. Sometimes, it seems, the author diverges from the notion of false friends, providing examples of words that are not false friends. It is a useful material, but not so much in the case of semantic false friends between French and Latvian.

More than thirty years have passed since their compilation. It seems that their revision could be useful to remove the words that are not false friends anymore, add a new sense to the words undergoing diachronic changes, and include newly emerged false friend examples. Such updated lexicographical material could be integrated into digital dictionaries to improve the information available on false friends. Since false friend dictionaries are in comparison very small and would therefore have a very limited target audience, it would not suffice to leave them as a stand-alone lexicographical material.

5. Conclusions

This paper sheds light on the treatment of false friends in the most recent English-Latvian and French-Latvian general dictionaries. Analysis revealed recurring patterns of management of false friends across several dictionaries. False friends in them are not specifically marked; labels are most frequently used to indicate differences in word use. In general, the treatment of false friends is sufficient if the provided meanings are clear and unambiguous. Yet, in some cases, only one equivalent was provided, thereby sustaining the illusion of complete equivalence. This is not helpful since diverging contextual, cultural and semantic references are not disambiguated.

The possibility of false friend marking was discussed to understand the potential and possibilities of such a solution. It is not being practised in the analysed dictionaries which might be due to space constraints – a significant limitation for printed dictionaries.

Existing false friend dictionaries could be revised to update entries where false friends have become “true friends” due to diachronic changes. Since the target audience for such dictionaries is minimal, it would be best to incorporate them into existing digital dictionaries, which would increase the likelihood of this information reaching a larger public. The above observations would be valid for lexicographic treatment of any language pairs with false friends.

References

Baldunčiks, J. (2006). Angļu-latviešu tulkotāja viltusdraugi: teorija un prakse. *Latvijas Universitātes raksti* 700, 260–269.

- Bankavs, A. (1989). *Les faux amis du traducteur franco-lettons*. Riga: P. Stučka LVU.
- Broz, V. (2008). Diachronic investigations of false friends. *Suvremena lingvistika* 34(66), 199–222.
- Corréard, M.-H. (2002). Are space-saving strategies relevant in electronic dictionaries? In A. Braasch, & C. Povlsen (Eds.), *Proceedings of the 10th EURALEX International Congress* (pp. 463–470). Center for Sprogteknologi.
- Gorbahn, A., & Hausmann, F. J. (1989). The dictionary of false friends. In F. J. Hausmann, O. Reichmann, H. E. Wiegand, & L. Zgusta (Eds.), *Wörterbücher, Dictionaries, Dictionnaires. An International Encyclopedia of Lexicography* Vol. 3 (pp. 2882–2888). Walter de Gruyter.
- Gouws, R. H., Prinsloo, D. J., & de Schryver, G.-M. (2004). Friends will be friends – true or false. Lexicographic approaches to the treatment of false friends. In G. Williams, & S. Vessier (Eds.), *Proceedings of the 11th EURALEX International Congress* (pp. 797–806). Université de Bretagne-Sud, Faculté des lettres et des sciences humaines.
- Granger, S., & Swallow, H. (1988). False friends: a kaleidoscope of translation difficulties. *Langage et l'Homme* 23/2, 108–120.
- Hartmann, R. R. K., & James, G. (1998). *Dictionary of Lexicography*. Routledge.
- Sabino, M. A. (2016). False cognates and deceptive cognates: issues to build special dictionaries. In T. Margalitadze, & G. Meladze (Eds.), *Proceedings of the 17th EURALEX International Congress* (pp. 746–755). Ivane Javakhishvili Tbilisi University Press.
- Veisbergs, A. (1994). *Latviešu-angļu, angļu latviešu viltus draugu vārdnīca*. Second edition. SI.
- Veisbergs, A. (1996). False friend dictionaries: a tool for translators or learners or both. In M. Gellerstam, J. Järborg, S.-G. Malmgren, K. Norén, L. Rogström, & C. R. Papmehl (Eds.), *Proceedings of the 7th EURALEX International Congress* (pp. 627–634). Novum Grafiska AB.
- Veldi, E. (2006). English-Estonian Dictionary of False Friends: Why was this Dictionary Needed? In A. Veisbergs (Ed.), *Pragmatic Aspects of Translation: Proceedings of the Fourth Riga International Symposium* (pp. 169–181). SIA JUMI.

Dictionaries:

ENG 1: *Angļu-latviešu vārdnīca*. (2000). Jāņa Sēta.

ENG 2: *Angļu-latviešu vārdnīca*. (2002). Avots.

ENG 3: *Angļu-latviešu vārdnīca*. (2007). Avots.

FR 1: *Franču-latviešu vārdnīca*. (2005). 2nd edition. Avots.

FR 2: *Franču-latviešu vārdnīca*. (2012). 2nd edition. Zvaigzne.

Websites:

ENG 4/FR 3: *Tildes vārdnīca*. Retrieved May 29, 2024, from <https://translate.tilde.com/#/>

Larousse: *Larousse French Dictionary*. Retrieved May 29, 2024, from <https://www.larousse.fr/dictionnaires/francais>

LVK2022: *Līdzsvarotais mūsdienu latviešu valodas tekstu korpuss*. Retrieved May 20, 2024, from <https://repository.clarin.lv/repository/xmlui/handle/20.500.12574/84>

MLVV: *Mūsdienu latviešu valodas vārdnīca*. Retrieved May 29, 2024, from <https://mlvv.tezaurs.lv/>

OLD: *Oxford Learner's Dictionaries*. Retrieved March 25, 2024 <https://www.oxfordlearnersdictionaries.com/>

Contact information

Evelīna Ķiršakmene
University of Latvia
evelina.kirsakmene@gmail.com

Daria Lazić

LEXICOGRAPHIC TREATMENT OF VOCABULARY RELATED TO AGE

The Example of Croatian and Danish

Abstract The paper deals with vocabulary related to age and age groups, its semantic and pragmatic characteristics, and challenges it can pose for lexicographic description. The inventory and treatment of lexical items from the domain of age in contemporary Croatian general dictionaries and in the most comprehensive contemporary Danish dictionary are analysed regarding the semantic and pragmatic information provided, as well as the presence of age stereotypes or strategies to mitigate them. In Croatian dictionaries, many terms denoting older people or phenomena relevant to them are not listed. Semantic and pragmatic information, such as connotations or appropriateness, is incomplete and inconsistent, and examples reflect common stereotypes of young and old people. In the Danish dictionary, the inventory of lexical items is more comprehensive, the descriptions are more detailed, and pragmatic information is commonly provided. The examples are mostly realistic and neutral. However, a comparison with entries related to some other social groups indicates a slightly higher tolerance for age-related stereotypes.

Keywords critical lexicography; social stereotypes; age; Croatian; Danish

1. Introduction

The fact that dictionaries are inseparable from the socio-cultural context in which they are compiled is especially pronounced when it comes to vocabulary related to social groups. Nowadays, following an increase in social sensitivity, dictionaries are often criticized for stereotypical or negative portrayals of social groups, which has incited research interest in the matter. Moreover, the meaning and use of lexical items related to social groups are usually tightly connected to the socio-cultural context. To facilitate proper understanding and use of these items, in addition to describing their denotative meaning, other information might be considered, e.g., pragmatic information such as appropriateness and situational constraints. Potentially offensive content, such as offensive terms or stereotypes that are contained in the typical context of use, is a common issue when dealing with vocabulary related to social groups (Cloete, 2014; Schutz, 2002). The problem is emphasized by the fact that contemporary dictionaries are commonly corpus-based. Corpora can offer valuable information on contexts in which words are used and aspects of meaning stemming from them (e.g., McEnery & Hardie, 2011), but the language use attested in corpora often reflects negative attitudes and stereotypes which exist in society.

Numerous studies on the lexicographic representation of social groups have been conducted within English lexicography, and they have regarded various social

groups and bases of discrimination, e.g., gender (Fournier & Russel, 1992; Moon, 2014), ethnicity and age (Moon, 2014), illness and disability (Norri, 2019). In Croatian and Nordic contexts, studies have mainly covered gender stereotypes (Fjeld, 2015; Pišković, 2017) and editorial practices regarding social sensitivity in particular dictionary projects (e.g., Jensen et al., 2018; Petersson & Sköldberg, 2020; Trap-Jensen, 2020 for Danish and Swedish). The treatment of lexical items denoting older people in Swedish dictionaries was analysed in Nikula (2008), while in the context of Croatian and Danish lexicography, the representation of age groups in dictionaries has not been investigated.

In the present paper, the inventory and treatment of lexical items related to age and age groups in Croatian contemporary general dictionaries and in the Danish web dictionary *Den Danske Ordbog* (DDO; <https://ordnet.dk/ddo>), the most comprehensive dictionary of contemporary Danish, will be analysed regarding the semantic and pragmatic information provided and the presence of age stereotypes or strategies to mitigate them. The following Croatian dictionaries will be analysed: *Rječnik hrvatskoga jezika* edited by Jure Šonje (RHJ; Šonje, 2000), *Veliki rječnik hrvatskoga standardnog jezika* (VRH; Jojić, 2015), the most recent comprehensive general dictionary, and the online dictionary database *Hrvatski jezični portal* (HJP; <https://hjp.znanje.hr>).

Previous research (Lazić & Mihaljević, 2021) has shown that the representation of social groups, such as gender and ethnic groups, in contemporary Croatian dictionaries is often biased, and improvements could be made to render the descriptions more socially sensitive and inclusive. The comparison with the Danish dictionary DDO has been motivated by the fact that it is continuously expanded and updated, and its articles are revised for stereotypes and outdated perceptions of social groups (Jensen et al., 2018; Trap-Jensen, 2020). It can thus offer insight into alternative lexicographic solutions in that regard.

The purpose of the study is threefold: (1) to outline pragmatic and semantic characteristics of lexical items related to age and their implications for lexicographic description, (2) to identify shortcomings in the treatment of such vocabulary in the existing dictionaries, and (3) to propose guidelines and improvements for future lexicographic work. Regarding the last point, solutions to some issues discussed in the paper which were applied in *Mrežnik* (<https://rjecnik.hr/mreznik>), a new Croatian web dictionary, will be presented.

2. Age and Ageism in Society and Language

Age is one of the main criteria for categorising people, along with gender. Gender and age groups appear in all cultures, and they are – unlike some other groups, such as ethnic or religious – based on both biological and social factors. Age is associated with various milestones in our lives (e.g., legal age for driving, parenthood, retirement), and it can affect certain aspects of our lives such as friendship patterns, career, or interests (Schneider, 2005, pp. 466–467). Unlike gender and some other social groups, age is a continuous category. Our perception of age is based on segmenting the life

course into discrete age categories, or age groups. Examples of such groups are baby, toddler, child, teenager, young adult, middle-aged, older adult, and elderly (Swift et al., 2018, p. 2). However, age categories vary in number and age range within a particular category, and category boundaries are often not clear-cut. Age categorisation can also vary by context (e.g., healthcare, sports, family relations), culture, or over time. For example, Swift et al. (2018, p. 2) mention a new age group, centenarians, which has emerged because of extended life expectancy and a growing number of people living to the age of 100.

As with other social categories, certain stereotypes are associated with age groups. Attributes ascribed to group members are both descriptive (e.g., character traits) and prescriptive (e.g., expected behaviour). Some of them are motivated by maturation and biological ageing processes or social structures, while others are mere generalisations (Swift et al., 2018, p. 2). However, although ageing is inevitable, its course and effects can vary greatly, and groups determined based on chronological age measured in years since birth include individuals of very different physical and psychosocial statuses. As pointed out by Schneider (2005, p. 467), the group of people older than 60 is quite heterogeneous – people in their 60s are normally still working, and they are “at the point where energy, income, and wisdom are at their peak.” A decade later, people are usually retired, and health problems are more common. The most negative stereotypes are usually associated with very old people, those over 80.

Stereotypical beliefs can encompass negative, neutral, and positive characteristics. When it comes to age stereotypes, those about older people are most pronounced, given that old age is often viewed negatively as a period of decline in physical and mental abilities, and older people as a group that loses its role in society. However, although negative stereotypes about this social group prevail, some of them are positive (Chasteen, 2002). According to the Stereotype Content Model (Fiske et al., 2002), in which stereotypes are viewed along two dimensions: warmth and competence, perceptions of older people are usually evaluated positively for warmth and negatively for competence (Fiske et al., 2002; Swift et al., 2018). Regarding specific characteristics, some negative ideas about the abilities, personality, and physical condition of older people that appear in the research literature are reduced ability to acquire new skills, lack of creativity, reduced adaptability, poor technological skills, poor driving, sluggishness, despondency, alienation, vulnerability, grumpiness, and intolerance. Positive perceptions include, among others, kindness, understanding, self-motivation, productivity, and reliability (Swift et al., 2018).

Although less pronounced and discussed than stereotypes of older people, negative stereotypes of younger people also occur. For example, research in the American context¹ has shown that they are sometimes perceived as spoiled, lacking respect, inexperienced, and prone to adopting radically progressive views. Among other things, such perceptions might affect their financial and business situation and security.

¹ Friedman, Harvey A. “Kids these days”: Why youth-directed ageism is an issue for everyone. Retrieved May 15, 2024, from <https://publichealth.wustl.edu/kids-these-days-why-youth-directed-ageism-is-an-issue-for-everyone/>.

Stereotypes that exist within a certain community are reflected in language and expressed, shared, and maintained through communication. Stereotypical notions can be observed in the labels used to refer to a certain social category and in the characteristics and behaviours attributed to it (Beukeboom & Burgers, 2019). The relationship between the perception of (old) age and language has been the subject of several studies by Frank Nuessel (e.g., Nuessel, 1982; 1984; 2002). Nuessel (2002) discusses designations for older adults, vocabulary about older adults and ageing, metaphoric language, proverbial language, slogans, names, and forms of address, and concludes that the lexicon related to older adults is “overwhelmingly negative” (Nuessel, 2002, p. 25).

In languages, there is commonly a variety of expressions available to refer to age groups. This is especially pronounced with designations for older people; since old age can be regarded as a social and linguistic taboo (Saporta, 1991), various euphemisms (e.g., *golden age*, *senior citizen*) and dysphemisms (e.g., *old codger*, *fossil*) have emerged. Moreover, as old age is frequently discussed in negative contexts such as illness, dependency, or economic burden for society, even seemingly neutral labels can acquire negative connotations (cf. Nuessel, 1984, p. 18, the example of *elderly*). In addition to differences in style and connotations, labels can be general age-group designations (*over-60s*, *older citizens*), or regard additional characteristics (*pensioner*, *grandpa*, *hag*). Finally, labels can differ in linguistic form (e.g., *oldster*, *the aged*, *older adult*). The use of adjectival and descriptive labels together with general nouns such as *people*, *person* can be preferred over noun labels as a means of avoiding essentialism and promoting individual differences and diversity within the group (Beukeboom & Burgers, 2019, p. 16; APA1). In English, preferred general terms are adjective and noun combinations with the comparative *older*: *older people*, *older adults*, etc., or descriptive labels such as *persons 65 years and older* (APA2). The comparative *older* is also used in Croatian (*stariji*) and Danish (*ældre*) – alone or with nouns denoting people – as a neutral reference to this age group.

3. Vocabulary Related to Age in Dictionaries

The perceptions of age and age groups discussed in the previous section, and their reflection in language and communication, have implications for how vocabulary related to age is treated in dictionaries. The first issue to address is the selection of headwords which will be included in the dictionary. Additionally, given the different designations available, their semantics and pragmatics (e.g., appropriateness, function) should be clearly described. Finally, a strategy for handling potentially offensive content, such as language reflecting negative and stereotypical ideas about age groups, should be established. In what follows, the treatment of age-related vocabulary in selected dictionaries will be analysed with respect to: (1) selection of headwords, (2) definitions, (3) collocations and usage examples, and (4) pragmatic information.

3.1 Selection of Headwords

In contemporary Croatian dictionaries, some common designations for age groups are not listed, and information on neutral and preferred labels for older people cannot be

found. This might be because these designations are multi-word expressions, which are usually not included as headwords in Croatian dictionaries. However, they are not listed within entries of constituent words either.

The designation most commonly used in official contexts (such as healthcare, legal, and academic communication) is *starija osoba* 'older person'. A similar expression is *osoba starije (životne) dobi*, which literally translates to 'a person of older age (in life)'. In the analysed Croatian dictionaries, the comparative *stariji* 'older' is neither listed as an independent entry nor accounted for within the entry *star* 'old'.

Substantival labels *starac* 'old man, oldster' and *starica* 'old woman', which are listed in dictionaries, do not denote the general population of older adults defined as people over 60 or 65 years, but very old people exhibiting negative characteristics associated with ageing (except for the positive stereotype *wise*). It is evident from the definition of the word *starac* in VRH – *a man who is many years old, who is at an advanced age* ["muškarac koji ima mnogo godina, koji je u visokoj životnoj dobi"] – and the examples that are provided: *grey / decrepit / wise / ailing old man; This knight was an old man of 90 years*. ["sijedi / oronuo / mudri / bolujući starac; Ovaj je vitez bio starac od 90 godina."].

Moreover, many figurative and euphemistic expressions used to refer to old age without mentioning it directly are also in the form of multi-word expressions and are thus not listed as headwords in dictionaries. While some of them are included as subentries or examples in the related single-word entries, others are not recorded. For example, one of the senses of the headword *dob* 'age' is exemplified by the following collocations: *younger / middle / maturer / elderly / advanced / third age* ["mlađa / srednja / zrelija / staračka / poodmakla / treća dob"]. The expressions such as *starija dob* discussed above or *zlatna dob* 'golden age' are not listed. A similar euphemistic expression *zlatne godine* 'golden years', which is well-attested in names of institutions, associations, initiatives, media programmes, etc. aimed at older people, is listed as a subentry in VRH, but only the sense *years of abundance, accompanied by the growth and development of everything* ["godine u kojima vlada obilje, koje prati rast i razvoj svega"] is accounted for, and not the old age.

Finally, some concepts relevant to older people are insufficiently represented. An example is care/retirement home. The expression *starački dom* 'old people's home' appears as an example under the entries *starački* (a relational adjective derived from the noun *starac* 'old man') and *dom* 'home'. The expression is used colloquially, but as it contains the negatively connoted adjective *starački*, more neutral terms such as *dom za starije osobe* 'older people's home' are used in official contexts and names of such institutions. However, they are not listed in the dictionaries.

In the Danish dictionary DDO, both the adjective *gammel* 'old' and its comparative *ældre* 'older' are listed as separate entries. In both entries, adjectival (e.g., *gammel kone* 'old woman') and substantival (e.g., *de gamle* 'the old') usages are recorded. When referring to people, they are defined with an approximate age range; *gammel* 'old' is defined as *a person of a relatively high age (more than approximately 80 years)*

[“person med en forholdsvis høj alder (mere end ca. 80 år)”]; *ældre* ‘older’ is defined as *of an age of more than approximately 60 years; relatively old – about people* [“med en alder på mere end ca. 60 år; forholdsvis gammel – om menneske”]. For both adjectives, similar examples with nouns denoting people are provided (*ældre medborgere* ‘older citizens’, *ældre dame* ‘older lady/woman’, *ældre ægtepar* ‘older married couple’, *ældre herre* ‘older gentleman/man’; *gammel dame/kone* ‘old lady/woman’, *gammel mand/herre* ‘old man/gentleman’, *gamle mennesker* ‘old people’). Even though the adjective *gammel* denotes an older age group than its comparative *ældre*, similarly as in Croatian, the comparative is often used for older individuals regardless of age as a more polite term to avoid negative connotations of the adjective *old* (cf. Wien, 2006; Blaakilde, 2004). Although this is not explicitly commented on, compounds cross-referenced at the end of the article can signal usage in official contexts: *ældrebolig* ‘elder housing’, *ældreorganisation* ‘organisation for older people’, *ældrepleje* ‘eldercare’, *ældrevenlig* ‘elder-friendly/age-friendly’, etc. These examples also show that unlike the Croatian dictionaries, DDO contains many entries which relate to the experience and living conditions of older people, e.g., various types of accommodation: *plejehjem* ‘care home’, *dagcenter* ‘daycentre’, *ældrekollektiv* ‘senior collective’, *pensionistbolig* ‘retirement housing’. In addition, the *Danish Thesaurus* is integrated in the dictionary entries (Nimb et al., 2018), so a list of different words and expressions from the domain of old age can be found in one place. The expressions are hyperlinked to the corresponding articles in the dictionary if they exist.

3.2 Definitions

Phenomena can be defined in different ways and from various perspectives, thereby reflecting values and norms within a particular community or among some of its members. This is evident from an often-quoted example in the Croatian context: in a dictionary published in 2000 (RHJ), a man is defined as a *male person* [“muška osoba”], while a woman is defined as a *human being of the opposite sex to a man, capable of giving birth to children and taking major responsibility for their upbringing and education* [“ljudsko biće po spolu suprotno muškarcu, koje može rađati djecu i preuzeti glavnu brigu za uzgoj i odgoj djece”]. Here, a man is presented as an absolute category, while a woman is defined in relation to a man, with her role as a mother being central (cf. Bratanić, 2005).

The form and content of definitions are especially important when defining potentially sensitive lexical items, such as items related to social groups. For that reason, stereotypical and possibly offensive ideas are sometimes addressed in dictionaries aiming to be socially sensitive. This can be illustrated by the definition of the word *pigefornærmet* ‘offended as a girl’ from DDO: *the one that is offended or hurt by something someone has said or done, and reacts in a way that is perceived as characteristic of an immature teenage girl, for example, by refusing to talk about it or rolling one’s eyes. USAGE the usage of the word can be perceived as problematic because it builds on stereotypical notions of girls and women* [“som er fornærmet eller såret over noget en anden har sagt eller gjort, og reagerer på en måde der opfattes som karakteristisk for en umoden teenagepige, fx ved at nægte at tale om det eller himle

med øjnene. SPROGBRUG Brug af ordet kan opleves problematisk fordi det bygger på stereotype forestillinger om piger og kvinder”].

However, the definition of an expression based on the stereotypic view of old people as conservative, old-fashioned, and irrelevant in their reasoning – *gammelmandssnak* ‘old man’s talk’ – does not contain a similar comment on possible offensiveness or distancing expressions such as *perceived as characteristic*. The definition reads: *expressions of opinions that reflect resistance to change, and are based on bygone circumstances, values, etc.* [“meningstilkendegivelser der afspejler modstand mod forandringer, og som bygger på tidligere tiders forhold, værdier el. lign.”]. On the other hand, a comment on gender and age stereotypes is included in the description of the meaning and usage of the word *tante* ‘aunt’ in the sense of ‘older woman with a snarky or grumpy demeanour’ and the adjective *tantet* ‘aunt-like, auntly’ derived from it.

It should be noted that comments on underlying stereotypes are not used extensively in DDO and have so far been included only in a small number of entries reflecting gender, sexuality, and ethnic stereotypes (Sanni Nimb, Senior Editor of DDO, personal communication, July 31, 2024). Stereotypes are pervasive in language. While exposing stereotypes can be desirable, particularly for very offensive terms, it would be excessive to comment on them in every instance. When discussing various categories of sensitive items and their lexicographic treatment, Cloete (2014), following Hartevelt & Van Niekerk (1996), links the appropriate treatment of sensitive lexical items to the category they belong to since not all categories of items are equally sensitive. For sensitive lexical items indicating stigmatised physical or mental conditions and phenomena, the author concludes that generally speaking, “the sensitive items in this category, e.g., *hunchback*, *mongoloid* and *water head*, are not experienced as hurtful and demeaning as racist and sexist terms because they are used to a lesser extent to deliberately insult those involved. Reference plays a larger role here than typification and insult. However, within the category, items can vary in their degree of sensitivity.” The differences in the lexicographic treatment of lexical items related to social groups discussed above might indicate that age stereotypes are not currently perceived as equally problematic as, for example, gender stereotypes.

Croatian dictionaries do not have a tradition of exposing and commenting on stereotypes (cf. Lazić & Mihaljević, 2021). For example, the idea that old women engage in useless activities has been lexicalised in several expressions containing the disparaging and/or colloquial word for an old woman – *baba* – or the adjective derived from it – *bapski*. Some of them are listed in the analysed Croatian dictionaries, such as *bapski lijek* ‘old woman’s remedy’, *bapske priče* ‘old woman’s tales’, and *bapska posla* ‘old woman’s business’. The stereotypes are not commented on, as is evident from the definition of the expression *bapski lijek* in VRH: *DEROGATORY home remedy that cannot be believed to help, a remedy that old women prepare and only they believe in* [“POGR domaći lijek za koji se ne može vjerovati da pomaže, lijek koji babe pripravlja i u koji samo one vjeruju”].

While many of the discussed expressions and their definitions reflect a negative view of ageing, it is interesting to compare the definition of the euphemistic expression

the third age in DDO and in VRH. In DDO, *den tredje alder* is defined as *the period of a person's life when they no longer work but can still be active* [“den periode af et menneskes liv, hvor det ikke længere arbejder, men stadig kan være aktivt”] and labelled as euphemistic. In VRH, it is simply defined as *a period of life after middle age, old age* [“doba života nakon srednjih godina, starost”], even though it is often used in affirmative contexts and is less negatively connoted than the word *starost* ‘old age’.

3.3 Collocations and Examples

Collocations and usage examples can reveal stereotypical ideas about social groups that exist in society and are communicated through language, e.g., characteristics and behaviours associated with members of a certain social category or the contexts in which they are commonly mentioned. Since modern dictionaries are usually corpus-based, the choice of collocations and examples will be influenced by corpus data.

The analysed Croatian dictionaries are either printed or based on printed dictionaries, thus containing only collocations and shorter examples. Consistent with the fact that neutral designations for older people are not listed in the dictionaries, collocations and examples mostly reflect negative aspects of ageing and undesirable traits associated with old people. As mentioned above, in VRH, adjectival collocates given for the noun *starac* ‘old man’ indicate physical decay (*oronuli* ‘decrepit’, *bolujuć* ‘ailing’), alongside the common positive stereotype *mudri* ‘wise’. Another adjective derived from the disparaging/colloquial word *baba* ‘old woman’ – *bablji* – is illustrated by collocates mostly denoting undesirable or low-value activities and behaviours such as *brbljanje* ‘babbling’, *kuhanje* ‘cooking’, *bajke* ‘fairy tales’, *slutnje* ‘forebodings’, *tlapnje* ‘fantasies’, *jezik* ‘tongue’, *posao* ‘business’. More positively connoted expressions such as *the third age* are not equipped with examples. Collocates and examples listed for expressions denoting young people also align with common stereotypes such as physical beauty and strength on one side and immaturity, naivety, and arrogance on the other, e.g., *mladić* ‘young man’: *darovit* ‘gifted’, *naočit* ‘handsome’, *neiskus* ‘inexperienced’, *snažan* ‘strong’; *mladički* ‘related to young men’: *drskost* ‘arrogance’, *obijest* ‘ingratitude’; *mladenački* ‘related to young people’ (adjective): *ljepota* ‘beauty’, *strast* ‘passion’; *mladenački* ‘like a young person’ (adverb): *naivan* ‘naive’.

In DDO, senses are exemplified by both shorter examples (collocations and phrases) and citations from corpora. Unlike the Croatian dictionaries, for items denoting age groups, collocations are mostly not provided, or they are not qualifying in nature, so they do not reflect stereotypes. An exception is the adjective *ældre* ‘older’, referring to the age of 60 or older, which is illustrated by collocations like *plejekrævende ældre* ‘older people requiring care’, *ældre og gangbesværede* ‘older and walking-impaired’, and *ældre og svagelige* ‘older and infirm’. These characteristics are not typical for the whole population over 60, but probably reflect the contexts in which older people are frequently mentioned in corpora.

While citations should also reflect typical usage, the lexicographer has greater freedom to choose examples that do not unnecessarily portray social groups in a stereotypical, negative, or one-sided manner. Examples related to old age in DDO seem to be realistic; some negative aspects of ageing are represented, for example, *gammel* 'old': *When a person grows old, their skin wrinkles, and the thought is not as powerful as when they were young* ["Når mennesket bliver gammelt, rynkes dets hud, og tanken har ikke den samme kraft som da det var ungt"], as well as positive aspects, e.g., *pensionist* 'pensioner': *As a pensioner, I myself decide how the day should be put together* ["Som pensionist bestemmer jeg selv, hvordan dagen skal skrues sammen"], or neutral, e.g., *ældre* 'older': *In Hanstholm, older people can decide for themselves whether they want to stay at home, or whether they want to move into one of the Care Centre's elder homes* ["I Hanstholm kan de ældre selv bestemme, om de vil blive hjemme, eller om de vil flytte ind i en af Plejecentrets ældreboliger"].

3.4 Pragmatic Information

Pragmatic information – information on the usage of lexical items – can cover a range of details related to, among other things, the communicative situation, linguistic context, or extralinguistic (encyclopaedic) knowledge. Pragmatic information facilitates the correct usage of lexical items and can thus be useful to users. Nonetheless, such information is scarce in Croatian monolingual dictionaries (Pritchard, 1999). Pragmatic information can be contained in different elements of lexicographic description: labels, definitions, usage notes, or examples. In this regard, the following issues will be discussed: (1) lexical items denoting people of a certain age being used as terms of address and (2) affective and attitudinal aspects of such items.

In DDO, the definition of the word *bessesar* 'grandpa' reads: *used jokingly or derogatorily when addressing or referring to an older man* ["bruges spøgende eller nedsættende ved tiltale til eller omtale af en ældre mand"]. The word *bessemor* 'grandma' is defined in a similar way. The usage of words for grandparents when addressing older people one is not related to is attested even in Croatian. Though such usage can convey humour or derogation, it is not always the case. For example, it is commonly used by healthcare workers.² Even though the intent behind such usage is not necessarily to denigrate the person, but rather to appear warm and welcoming, it is patronising in nature and perceived by many as offensive. The usage is attested in corpora – e.g., in the following example from the Croatian web corpus *hrWaC*: *When I got her off the road, I (...) asked her: 'so granny, where did you go?'* ["Kad sam je maknula sa ceste (...) pitam je: 'pa bako kud ste vi krenuli??'"]. However, while the usage of words for grandparents when referring to old people is mentioned in some of the analysed Croatian dictionaries, none of them describes the usage of these words as forms of address.

Similarly, in both languages, there are expressions denigrating people because of their young age and associated traits such as immaturity, lack of experience, or arrogance. Some of them are commonly used when addressing people, as is evident

² Tešija, Tijana. *Oslovljavanje starijih osoba*. Retrieved May 18, 2024, from <https://www.seniori.hr/stranice/psihosocijalne-teme/1511-oslovljavanje-starijih-osoba>.

from these examples: *grønskolling*: young, inexperienced (and timid) person USAGE derogatory; especially as a form of address [“ung, uerfaren (og frygtssom) person SPROGBRUG nedsættende; især i tiltale”]; *snothvalp*: young, immature (and arrogant) person USAGE derogatory; especially as a form of address [“ung, umoden (og fræk) person SPROGBRUG nedsættende; især i tiltale”]. The similar Croatian word *balavac* is defined in VRH as *young, immature, and insufficiently experienced person* [“mlada, nezrela i nedorasla osoba”]. It is labelled as derogatory, but no further information on its usage is provided, even though it can also be used when addressing people. Usage of the adjective *mali* ‘little’ as a form of address is recorded in HJP and VRH, but while no information on the attitude of the speaker or effect on the addressee is provided in HJP, in VRH, it is labelled as familiar. In DDO, usage of the equivalent adjective *lille* when addressing people is described as *used as affectionate, confidential, or possibly condescending term, often as a form of address* [“bruges som kærlig, fortrolig eller evt. nedladende betegnelse, ofte i tiltale”], which indicates that such usage is not necessarily intended or perceived as positive. Similar effects are mentioned for other words denoting young age or small size, especially when directed at adults. The entry *bitte* ‘very little, tiny’ contains the remark that it is used when addressing or referring to a child in an affectionate way, but it can be humorous or ironically derogatory when used to address or refer to an adult.

3.5 Possible Solutions

Dictionaries should reflect the way language is used at a certain time without falsifying reality. However, some lexicographic choices and strategies can have an impact on representation of social groups in dictionaries. When describing vocabulary related to age groups (and social groups in general) in the new Croatian web dictionary *Mrežnik* (<http://ihjj.hr/mreznik/>), an initiative has been taken to overcome some of the shortcomings observed in existing dictionaries. Unlike the analysed Croatian dictionaries, *Mrežnik* is born-digital, so there is no space limitation. The use of words is illustrated by a list of collocates and examples extracted from corpora, and usage notes can be added to explain the use and meaning of words in more detail when necessary.

Among other things, the following information regarding the vocabulary related to age groups has been included:

- various designations for older people, such as the comparative *stariji* ‘older’, expressions *starija osoba* ‘older person’, and *osoba starije životne dobi* ‘person of an older age in life’, references to old age such as *treća dob* ‘third age’, *četvrta dob* ‘fourth age’, *zlatna dob* ‘golden age’;
- usage notes containing information on neutral designations for older people, differences between similar expressions in meaning, style, and contexts of use, appropriateness of certain expressions, etc.; e.g., addressing old people with words denoting grandparents is explained as follows: “In informal situations, old people are sometimes addressed as *baka* ‘grandma’ and *dida/deda* ‘grandpa’. Although speakers often use these expressions with good intentions, for example to achieve intimacy with the interlocutor, they can be

perceived as a lack of respect. Expressions *gospođa* ‘Madam’ and *gospodin* ‘Sir’ should be used as a polite way of addressing adult people regardless of age”;

- collocates in *Mrežnik* are based on corpus data; they can reflect stereotypes and negative views on ageing, for example, collocates of the expression *starija osoba* in the Croatian web corpus *hrWaC* (Ljubešić & Klubička, 2016) mostly belong to the domains of illness, care, and other vulnerable groups (e.g., *demented, immobile, lonely older people; older and infirm people, older people and children*), and it will be mirrored in the dictionary;
- examples should be neutral and illustrate different contexts in which an item is used e.g., the following examples have been chosen for the expression *starija osoba*;
- *The Croatian Ministry of Health and Social Welfare’s reference centre for the health protection of older people has issued a risk warning for the older people in Croatia due to low air temperatures.* [“Referentni centar Ministarstva zdravstva i socijalne skrbi Hrvatske za zaštitu zdravlja starijih osoba izradio je upozorenje rizika za starije osobe u Hrvatskoj radi niskih temperatura zraka.”];
- *If you are an older person and you move too little, start with regular walking, so that you can get back in good shape.* [“Ako ste starija osoba i malo se krećete, započnite sa redovitim hodanjem, tako da se vratite u dobru kondiciju.”].

4. Conclusion

Given society’s negative view of ageing and old age, these phenomena are a kind of social taboo. It is reflected in the language, for example, in the existence of many euphemisms and dysphemisms denoting old people and old age. Moreover, since age is a continuous category, it can be segmented into age groups in different ways, and groups and their boundaries are context-dependent. Consequently, there are various expressions denoting age and age groups, and they can differ in connotations and appropriateness in certain contexts. Another issue which can have implications for the description of vocabulary related to age are stereotypes and attitudes which are either lexicalised in words and expressions or appear in their typical context of use. While the role of the dictionary is to accurately mirror linguistic reality, the lexicographer might want to avoid unnecessary stereotyping and stigmatizing of social groups. Although negative stereotypes of old people are more prominent, they can also concern other age groups.

The analysis of Croatian contemporary dictionaries has revealed that many terms denoting older people or phenomena relevant to them are not listed in dictionaries. For the items that are included, information on semantic and pragmatic differences, such as connotations, register, or appropriateness in certain situations, is incomplete and inconsistent. Furthermore, examples and collocations in dictionaries maintain stereotypes of old people as weak, sick, dependent, etc., while young people are associated with physical beauty and strength on one side and immaturity, naivety, and arrogance on the other. In comparison, the inventory of lexical items related to age

and age groups in the Danish dictionary DDO is more comprehensive. Descriptions of lexical items related to age groups in DDO are more systematic and detailed, and pragmatic information is more commonly provided. For example, the use of certain items as terms of address and associated attitudes have been systematically described, while such information is missing in Croatian dictionaries. The examples and citations provided in DDO seem to be realistic and neutral with reference to both positive and negative aspects of ageing. However, a comparison with entries related to some other social groups indicates that there is a slightly higher tolerance for age-related stereotypes.

References

APA1: *General Principles for Reducing Bias – APA Style*. Retrieved May 20, 2024, from <https://apastyle.apa.org/style-grammar-guidelines/bias-free-language/general-principles>

APA2: *Age – APA Style*. Retrieved May 20, 2024, from <https://apastyle.apa.org/style-grammar-guidelines/bias-free-language/age>

Beukeboom, C. J., & Burgers, Ch. (2019). How Stereotypes Are Shared Through Language: A Review and Introduction of the Social Categories and Stereotypes Communication (SCSC) Framework. *Review of Communication Research*, 7, 1–37. doi.org/10.12840/issn.2255-4165.017

Blaakilde, A. L. (2004). Løber tiden fra Kronos? Om kronologiseringens betydning for forestillinger om alder. *Tidsskrift for kulturforskning*, 3(1), 67–85.

Bratanić, M. (2005). Mjesto žene u rječniku. In D. Stolac, N. Ivanetić, & B. Pritchard (Eds.), *Jezik u društvenoj interakciji* (pp. 37–46). Hrvatsko društvo za primijenjenu lingvistiku.

Chasteen, A. L., Schwarz, N., & Park, D. C. (2002). The Activation of Aging Stereotypes in Younger and Older Adults. *The Journals of Gerontology: Series B*, 57(6), P540–P547. doi.org/10.1093/geronb/57.6.P540

Cloete, A. E. (2014). The treatment of sensitive items in dictionaries. In R. H. Gouws et al. (Eds.), *Dictionaries. An International Encyclopedia of Lexicography. Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography* (pp. 482–486). De Gruyter Mouton.

DDO: *Den Danske Ordbog*. Retrieved May 25, 2024, from <https://ordnet.dk/ddo>

Fiske, S. T., Cuddy, A. J. C., Glick, P., & Xu, J. (2002). A model of (Often Mixed) Stereotype Content: Competence and Warmth Respectively Follow from Perceived Status and Competition. *Journal of Personality and Social Psychology*, 82(6), 878–902. doi.org/10.1037/0022-3514.82.6.878

Fjeld, R. V. (2015). Om ordbokseksempler og stereotypisering av kjønn i noen nordiske ordbøker. In C. Sandström et al. (Eds.), *Perspektiv på lexikografi, grammatik och språkpolitik i Norden* (pp. 35–65). Institutet för de inhemska språken.

Fournier, H. S., & Russell, D. W. (1992). A study of sex-role stereotyping in the Oxford English Dictionary 2E. *Comput Hum*, 26, 13–20.

Friedman, H. A. "Kids these days": Why youth-directed ageism is an issue for everyone. Retrieved May 15, 2024, from <https://publichealth.wustl.edu/kids-these-days-why-youth-directed-ageism-is-an-issue-for-everyone/>

Harteveld, P., & Van Niekerk, A. E. (1996). Policy for the Treatment of Insulting and Sensitive Lexical Items in the Woordeboek van die Afrikaanse Taal. In M. Gellerstam et al. (Eds.), *Proceedings of the 7th EURALEX International Congress* (pp. 381–393). Novum Grafiska AB.

HJP: Hrvatski jezični portal. Retrieved May 25, 2024, from <https://hjp.znanje.hr>

Jensen, J., Lorentzen, H., Nimb, S., Svendsen, M.-M. M., & Trap-Jensen, L. (2018). Thaipiger, muskelhunde og fulde svensker: nedsættende ord, stereotyper og ligestilling i Den Danske Ordbog. *Nordiske Studier i Leksikografi*, 14, 141–151.

Jojić Lj. et al. (Eds.). (2015). *Veliki rječnik hrvatskoga standardnog jezika*. Školska knjiga.

Lazić, D., & Mihaljević, A. (2021). Stereotypes and Taboo Words in Dictionaries from a Diachronic and a Synchronic Perspective – The Case Study of Croatian and Croatian Church Slavonic. In Z. Gavrilidou, M. Mitsiaki, & A. Fliatouras (Eds.), *Proceedings of XIX EURALEX Congress: Lexicography for Inclusion, Vol. II*. (pp. 101–111). Democritus University of Thrace.

Ljubešić, N., & Klubička, F. (2016). *Croatian web corpus – hrWaC 2.1*. Slovenian language resource repository CLARIN.SI, ISSN 2820-4042. Retrieved April 30, 2024, from <http://hdl.handle.net/11356/1064>

McEnery, T., & Hardie, A. (2011). *Corpus Linguistics: Method, Theory and Practice*. Cambridge University Press.

Moon, R. (2014). Meanings, Ideologies, and Learners' Dictionaries. In A. Abel, C. Vettori, & N. Ralli (Eds.), *Proceedings of the XVI EURALEX International Congress: The User in Focus* (pp. 85–105). Institute for Specialised Communication and Multilingualism.

Mrežnik: Hrvatski mrežni rječnik – Mrežnik. Retrieved May 25, 2024, from <http://ihjj.hr/mreznik/>

Nikula, K. (2008). Pensionär – aldrig i livet. De äldre i ordböckerna. *Nordiske studier i leksikografi*, 9, 337–351.

Nimb, S., Sørensen, N. H., & Troelsgård, T. (2018). From Standalone Thesaurus to Integrated Related Words in the Danish Dictionary. In J. Čibej et al. (Eds.), *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts* (pp. 916–923). Ljubljana University Press.

Norri, J. (2019). Treatment of words for illness and disability in monolingual English dictionaries. *International Journal of Lexicography*, 33(3), 227–250.

Nuessel, F. (1982). The language of ageism. *Gerontologist*, 22, 273–276.

Nuessel, F. (1984). Ageist Language. *Maledicta*, 22, 17–28.

Nuessel, F. (2002). Language about Ageing. In A. J. Ekerdt (Ed.), *Encyclopedia of Aging, Vol 3*. (pp. 757–762). Macmillan.

Petersson, S., & Sköldberg, E. (2020). To discriminate between discrimination and inclusion: a lexicographer's dilemma. In Z. Gavrilidou, M. Mitsiaki, & A. Fliatouras (Eds.), *Proceedings of XIX EURALEX Congress: Lexicography for Inclusion, Vol. I* (pp. 381–386). Democritus University of Thrace.

Pišković, T. (2017). Perpetuiranje rodni stereotipa u hrvatskim rječnicima. *Romanoslavica*, 52(2), 343–363.

Pritchard, B. (1999). O pragmalingvističkim informacijama u rječniku. In L. Badurina et al. (Eds.), *Teorija i mogućnosti primjene pragmalingvistike* (pp. 663–674). Hrvatsko društvo za primijenjenu lingvistik.

Saporta, S. (1991). Old Maid and Dirty Old Man: The Language of Ageism. *American Speech*, 66(3), 333–335.

Schneider, D. J. (2005). *The Psychology of Stereotyping*. Gullford Press.

Schutz, R. (2002). Indirect Offensive Language in Dictionaries. In A. Braasch, & C. Povlsen (Eds.), *Proceedings of the 10th EURALEX International Congress* (pp. 637–641). Center for Sprogteknologi.

Swift, H. J., Abrams, D., Drury, L., & Lamont, A. A. (2018). Categorization by Age. In T. K. Shackelford et al. (Eds.), *Encyclopedia of Evolutionary Psychological Science*. doi.org/10.1007/978-3-319-16999-6_2431-1

Šonje, J. (Ed.). (2000). *Rječnik hrvatskoga jezika*. Leksikografski zavod Miroslav Krleža – Školska knjiga.

Tešija, Tijana. *Oslovljavanje starijih osoba*. Retrieved May 18, 2024, from <https://www.seniori.hr/stranice/psihosocijalne-teme/1511-oslovljavanje-starijih-osoba>

Trap-Jensen, L. (2020). Inklusion eller mindretalsdiktatur? Om politisk korrekthet, minoritetshensyn og leksikografisk deskriptivisme i *Den Danske Ordbog*. *LexicoNordica*, 27, 137–156.

Wien, Ch. (2006). Hvornår blev de gamle ældre? *Gerontologi*, 22(3), 1–7.

Acknowledgements

This paper was written within the research project *Croatian Web Dictionary (Mrežnik) – Phase 2*, funded by the European Union – NextGenerationEU. Views and opinions expressed are however those of the author only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them.

Contact information

Daria Lazić

Institute for the Croatian Language

dlazic@ihjj.hr

Irina Lobzhanidze, Rusudan Gersamia, and Nino Tsulaia

COMPILING A BILINGUAL MEGRELIAN-ENGLISH ONLINE DICTIONARY Preserving Endangered Kartvelian Languages

Abstract The paper outlines one of the results of the project dedicated to one of the endangered Kartvelian languages, especially Megrelian. Providing data collection and documentation through fieldwork implemented in Samegrelo (Georgia), the project aims to comprehensively document the Megrelian language and encompasses the development of the annotated corpus, sketch grammar, and a bilingual dictionary. As a result, a bilingual Megrelian-English dictionary has been compiled using the Fieldwork Language Explorer (FLeX) and combining technological and traditional lexicographic approaches. We provide numerical examples to highlight the language structure and its application to the compilation of the dictionary, discussing its application to language preservation issues. The paper is subdivided into four parts: 1. Introduction, which outlines the project dedicated to the documentation of Megrelian language within the framework of the project financed by the Shota Rustaveli National Science Foundation (FR-21-993-3, 2021-2025); 2. Lexicographic insights on the Megrelian-English dictionary, which highlights the challenges of preserving endangered Megrelian language; 3. Macro- and micro-structures of the Megrelian-English dictionary, which emphasizes the structure of the dictionary compiled using FLeX and provides information on its licensing and accessibility; 4. Conclusions underscore the importance of this lexicographic effort and its application to the preserving of Megrelian language.

Keywords endangered languages; Kartvelian linguistics; Megrelian language

1. Introduction

The Kartvelian language family is characterized by its relatively uniform sound system, a well-developed system of word inflection and derivation, which involves a large variety of grammatical affixes and internal stem inflection and, the split ergativity of the sentence. Georgian, the most widely spoken Kartvelian language, has a rich literary tradition starting from the fifth century (Chikobava, 2008 [1952]; Shanidze, 1976; Sarjveladze, 1997, and others) and serves as the official language of Georgia. Svan, spoken in the mountainous regions of north-western Georgia, reveals significant phonological and morphological differences from other Kartvelian languages. Megrelian and Laz, spoken in western Georgia and north-eastern Turkey, respectively, share a close relationship and display remarkable similarities in terms of vocabulary and grammar. Despite the significant role played by the Kartvelian languages in shaping the cultural and national identity of the Georgian people, their present situation has undergone notable changes. Especially, Georgian, despite having a relatively small number of speakers, has managed to maintain its resilience. At the same time, other Kartvelian languages such as Svan, Megrelian, and Laz have unfortunately been listed as “increasingly endangered” in the UNESCO Atlas of the

World's Languages in Danger (2021). Svan, Megrelian and Laz languages transmit unique cultural knowledge encoding oral tradition that is under the threat of disappearance. The biggest challenge to preserve these unique languages is associated not only with the importance of their documentation, but also with the preparation of a dictionary using the data collected during the fieldworks.

In any case, preserving an endangered Megrelian language presents a lot of challenges provoked by linguistic and sociocultural aspects, including:

- **Scarcity of resources:** one of the most important challenges is the limited availability of linguistic resources, especially dictionaries. From the contemporary point of view, Megrelian faces a shortage of materials, especially written textual or other media, documented contemporary data and linguistic studies of the above-mentioned data using modern technologies. The absence of a contemporary appropriately annotated Megrelian corpus affects understanding of various linguistic aspects of grammatical structure of Megrelian and its vocabulary. The only one resource available by this moment is a corpus of Megrelian (GNC Megrelian) (Gippert et al., 2011–2024), which includes a small quantity of texts consisting of 89 404 words with punctuation and documented in the previous century without linguistic annotation. All the above-mentioned facts complicates the compilation of grammars, textbooks and dictionaries. As a result, in addition to very old notes on Megrelian like wordlists mentioned in (Gippert, 1992; 2016), the dictionaries of Megrelian are subdivided into printed dictionaries published in the past century (Kipshidze, 1914; Charaia, 1997; Eliava, 1999; Kajaia, 2000–2009, and others) and online dictionaries created during the last time and based on printed dictionaries (Kajaia, 2000–2009; Kobalia, 2010–2020, and others). These dictionaries provide valuable insights about the Megrelian lexicography, but do not capture the contemporary linguistic situation.
- **Urgency of the Task:** With each passing generation, the number of proficient speakers decreases, making the urgency of the preservation task even more complicated and important. The younger generations do not sufficiently acquire the endangered Megrelian language due to societal and educational influences, leading to a significant gap in generational transmission between the linguistic heritage of older generations and the linguistic proficiency of the younger population and implies that data mentioned above and collected a century ago not only fails to represent the current grammatical structure of a language but also inadequately reflects the present condition of its dictionary.
- **Globalisation and Georgian language influence:** Increased globalisation and Georgian language influence led to the adoption of the Georgian language and lifestyles, diminishing the value placed on preserving the Megrelian language and leading to a decline in its usage and significance. Addressing this issue requires not only an acknowledgment of the broader sociocultural dynamics influencing language choices but also the implementation of linguistic revitalization efforts, including the creation and compilation of resources (online corpora, grammars, dictionaries, etc.).

As a result, a compilation of Megrelian-English dictionary can be considered as an attempt in slowing down the negative impacts on endangered Megrelian making it more accessible worldwide. Its compilation became possible in parallel with the compilation of the annotated online corpus of Megrelian consisting of 97,393 tokens (60,783 types) collected during the language documentation project financed by the Shota Rustaveli National Science Foundation (FR-21-993-3, 2021-2025). The main scope of the project was to collect contemporary data via fieldwork, to process them using the Fieldwork Language Explorer (FLeX) and to compile a corpus, sketch grammar and an online dictionary combining technological and traditional lexicographic approaches. The dictionary is available on the official site of the Corpus of Megrelian language (<https://xmf.iliauni.edu.ge/>). The resource is intended to serve not only as a comprehensive reference of the contemporary linguistic condition of Megrelian language, but also as a specialized resource for those interested in revitalization of Megrelian language.

2. Language Structure and Lexicographic Insights on the Megrelian-English Dictionary

The compilation of linguistic specification is always closely connected to the macro- and the micro- structures of a dictionary. Deciding on the types of entry the dictionary will include and organizing the headword list are macrostructure decisions, but planning the entries in the dictionary and deciding on their structure and components are microstructure decisions (Atkins & Rundell, 2008).

A single entry can consist of a headword with its accompanying different meanings and, appropriately, it should be stored in the dictionary database (DB) in an alphabetical order. In the case of European normalized languages, the problem of alphabetization cannot be considered as a challenge. In the case of the Megrelian language, the headword list as well as the structure of entries in the dictionary are always under the impact of complex morphological structure, huge amount of word and/or form formation affixes and affect the compilation of the dictionary as a whole.

Thus, the issues under consideration are subdivided into three main parts: nominal and verbal inflections of Megrelian, which cause lemmatization and alphabetization problems, and, correspondingly, the access by end-users to the headwords of dictionary entries.

2.1 Nominal Inflection

In Megrelian, the structures of nouns, adjectives, numerals and pronouns have something in common, but the quantity of slots and formation models are different. Generally, the formation of nominal inflection is carried out by suffixation; only the degrees of an adjective are formed by means of circumfixation.

Types of stems are a base for different types of inflections; therefore, nouns are

subdivided into common and proper nouns with different types of inflections. Pronouns are subdivided into personal, demonstrative, indefinite, possessive, interrogative, relative, reciprocal, negative, determinative and reflexive ones; numerals into cardinal, ordinal fractional ones; and adjectives into gradable and non-gradable ones, i.e., those which can produce degrees of comparison and those which cannot. The main morphological features which affect the formation of nominal inflection are as follows: case (nominative, ergative, dative, genitive, instrumental, adverbial and vocative), number (singular and plural), postpositions associated with concrete cases and clitics. Additional categories are degree for adjectives and person for pronouns. The schemes of nominal formation are as follows:

- Noun: Stem -> Consonant epenthesis -> Number -> Vowel Epenthesis -> Case -> Emphatic Vowel -> Postposition -> Focus -> Emphatic vowel -> Particles[1, 2, 3] -> Conjunction
- Numeral: Stem -> Number -> Vowel Epenthesis -> Case -> Emphatic Vowel -> Postposition -> Focus -> Emphatic vowel -> Particles[1, 2, 3] -> Conjunction
- Pronoun: Stem -> Number -> Vowel Epenthesis -> Case -> Postposition -> Focus -> Emphatic vowel -> Particles[1, 2, 3] -> Conjunction
- Adjective: Degree -> Stem -> Degree -> Consonant epenthesis -> Number -> Vowel epenthesis -> Case -> Emphatic vowel -> Postposition -> Focus -> Particle[1, 2, 3] -> Conjunction.

Using the above-mentioned schemes, it becomes clear that the quantity of slots depends on the part of speech (PoS). Thus, the possible quantity of slots in nominals varies between 11 and 13 slots (1–4).

- (1) *tʃxu-l-ɛp-ifɔ-t-i-a-va-ɔ*
COW-E-PL-BEN-POST-EMPH-QUOT3-QUOT3-QUOT
'As it was said, and for cows'
- (2) *tʃitʃɛ-l-ɛp-ifɔ-t-i-a-va-ni*
little-E-PL-BEN-POST-EMPH-QUOT3-QUOT3-that
'As it was said, and for little ones that'
- (3) *vitɔxut-ifɔ-t-i-a-va-ɔ*
fifteen-E-PL-BEN-POST-EMPH-QUOT3-QUOT3-QUOT
'As it was said, and for fifteen'
- (4) a) *atena-Ø*
this-SG.NOM
'This'
b) *atɛn-ɛp-ifɔ-t-i-a-va-ɔ*
this.PROX-PL-BEN-POST-EMPH-QUOT3-QUOT3-QUOT
'As it was said, and for them'

Thus, from the lexicographic point of view, the primary wordform of a dictionary entry is the lemmatized form of a word. In case of nominals, the lemmatized form is a form in the nominative singular, which makes it easy to find nominals by typing the initial letters of a stem. This rule works for adjectives as well, because they are listed in their positive degree.

2.2 Verbal Inflection

In Megrelian, like in other Kartvelian languages, the main morphological features which affect the formation of verbal inflection are as follows: the TAM (tense-aspect-mood) series, which specify case-marking and linking between participants like agent and patient by means of preverbs, version vowels and thematic suffixes, voice subdivided active, medium and passive, personality, which covers unipersonal, bipersonal and tripersonal verbs, and number. A Megrelian verb contains many morphemes, which from one point of view are typical for agglutinating structures, but from another point of view are characteristic of inflected ones. The Megrelian verbal paradigm can be considered as a mixed one. The scheme of verbal formation is as follows:

Negation particle-> Affirmative particle -> Preverb -> Aspect -> Evidentiality -> Subject&Object agreement -> Applicatives, voice, causation or potentialis -> Stem -> Augment -> Voice, causation -> Thematic suffix or potentialis -> Tense&Aspect -> Subject&Object agreement -> Emphatic vowel -> Mood

Most verbs have preverbs lexically associated with them, although there is a group of verbs that do not have preverbs. Preverbs can be classified as a closed class always associated with verbs (5).

(5)	a) <i>gɔnɔ-rt-u</i> prv-go-s3SG(AOR) 'moved, transferred'	b) <i>mik'i-rt-u</i> PRV-turn-s3SG(AOR) 'turned around'	c) <i>ak'ɔ-rt-u</i> PRV-turn-s3SG(AOR) 'rushed in, burst in'

In addition, any kind of morpheme gives information about the grammatical function of the word. A Megrelian verb, generally, uses bound morphemes to show its grammatical function. The main types of use are as follows (6–7):

- (6) affixation, e.g.,
- a) *mɔ-tmɔ-v-ɔ-zim-ap-u-an-d-i-t-i-k'ɔ-ni*
PRV-IMPFV-S1-CAUS-measure-CAUS-AUG-TS-IMPF-PM1/2-PL1/2-COND-CONJ
'if I measured smth.'
- b) *gɛmuan-i-ɛ-n-i-a-ni*
delicious-NOM-BE-STAT-S3-EMPH-QUOT3-THAT
'This is delicious'
- (7) root alternation, e.g.,
- a) *sk'id-u-n-ɔ?*
stay-PASS-S3SG-QUOT(PRS)
'Is he/she/it staying?'
- b) *i-t'-ɛn-s*
APPL.SUBJ-stay-TS-S3SG(PRS)
'He/She/It is leaving'

The stemming of the verbal form is closely connected to the generation for subject- and object-based paradigms (8–9) and provision of their further analysis.

- (8) *v-ɔ-rts'q'-ε-k*
SUBJ1-APPL.INDIR-see-STAT-SUBJ1SG(PRS)
'sees'
- (9) *m-ɔ-rts'q'-ε-k*
OBJ1-APPL.INDIR-see-STAT-SUBJ1SG(PRS)
'view, observation'

Keeping in mind that the majority of the existing Megrelian dictionaries follow a mixed type of verbal presentation, which covers not only verbal nouns and adjectives (10), but also the third person singular forms (11), both of these forms are represented in the dictionary.

(10)	a) <i>pun-ap-a</i> receive-TS-SG.NOM 'reception'	<i>pun-ap-il-i</i> receive-TS-PTCP-NOM(PST.PASS) 'received'
	b) <i>ε-tʃɔp-u-a</i> PRV-take-OBLV-SG.NOM 'taking'	<i>ε-ma-tʃɔp-al-i</i> PRV-PTCP-take-PTCP-NOM(ACT) 'he/she/it who takes smth.'
(11)	<i>ɔ-pun-u-an-s</i> APPL.INDIR-boil-AUG-TS-S3SG(PRS) 'boiling'	<i>ε-tm-ε-tʃɔp-un-s</i> PRV-PRFV-boil-TS-S3SG(PRS) 'boiled'

From the lexicographic point of view, the problems of lemmatization as well as alphabetization are closely connected to the existence of particles, preverbs, personal markers, applicatives, voice, markers of evidentiality, potentialis and causation, which affect the initial point of a verbal entry and change the lexical meaning of the verb. Personal markers, applicatives and voice markers are generally used to indicate agent, patient and beneficiary and, respectively, subject, direct and indirect object relations, but they do not change the lexical meaning of a verb.

2.3 Lemmatization and Alphabetization

2.3.1 Lemmatization

Lemmatization is the process of deriving the base form or lemma of a word from one of its inflected forms and, generally, refers to the language vocabulary. For Kartvelian languages, including Megrelian, lemmatization can be considered a quite difficult task. This difficulty arises because of a huge quantity of agglutinating affixes. For example, a verbal scheme may have at least 8 prefixes associated with separate linguistic features and preceding the root. As a result, the determination of the lemma (i.e., initial point of a word) is a very complicated task.

Additionally, the lemmatization is a part of the linguistic normalization process, which is very complicated in case of endangered low-resourced languages. The linguistic normalization can be thresholds: morphological, syntactical and lexicosemantically. For the lexicographic purposes, the most important is morphological normalization, which strictly follows the Morpho-syntactic Annotation Framework (MAF). Following MAF, the morphological normalization performed by means of lemmatization reduces verbal forms to the infinitive, nominal forms to the nominative singular, comparative and superlative of adjectives to the positive. The only exception described in MAF is a verbal lemma in Arabic represented in the third person singular with the accomplished aspect (ISO/DIS 24611, 2012).

The MAF framework adopted for Modern Georgian can be considered as useful for other Kartvelian languages as well including Megrelian. In case of Megrelian, it means that the lemma for the nominal paradigm is represented with the nominative singular (12), for the adjectives – the prevalence is given to the positive degree (13), while for the verbal paradigm – the lemma can be represented in the form of masdar, while the lexical entry can contain the third person singular in present or future indicative (14) as well.

(12)	a) <i>k'ɔtʃ-i</i> man-SG.NOM 'man'	b) <i>dida-Ø</i> mother-SG.NOM 'mother'		
(13)	a) <i>did-i</i> big-SG.NOM 'big'	b) <i>mɔ-did-ɛ-Ø</i> DIM-big-DIM-SG.NOM 'bigger'	c) <i>ma-did-a-Ø</i> EQT-big-EQT-SG.NOM 'big like something'	d) <i>u-did-aʃ-i</i> SUP-big-SUP-SG.NOM 'the biggest'
(14)	a) <i>tʃk'ɔm-u-a</i> eat-OBLV-SG.NOM 'eating'	b) <i>ɔ-tʃk'ɔm-u</i> PRV-eat-TS-S3SG(AOR) 'He/she/it ate'	c) <i>tʃk'ɔm-un-s</i> eat-TS-S3SG(PRS) 'he/she/it eats'	d) <i>ɔ-tʃk'ɔm-un-s</i> PRV-eat-TS-S3SG(FUT) 'he/she/it will eat'

All of the above-mentioned approaches affect the compilation of Megrelian-English dictionary.

2.3.2 Alphabetization

The alphabetization problem already described concerning Modern Georgian language (Lobzhanidze, 2018) is relevant for the Megrelian language as well. Following Atkins & Rundell (2008), the problem of alphabetization is relevant to print dictionaries, but not to electronic ones, and it does not exist if all the headwords are single words, but does arise if the headword list contains MWEs. In Megrelian language, like in other Kartvelian languages, the problem arises concerning the verbal forms. In spite of the lemma used in the form of masdar (14), the verbal stem in the verbal template occupies the ninth slot and it makes impossible to arrange verbs in alphabetical order without paying attention to preverbs, prefixal person markers, applicatives etc. As a result, in the case of Megrelian, the alphabetization problem affects not only print but electronic dictionaries as well. In order to allow the user to access the headword of a dictionary entry, the headwords have to be arranged, from one

point of view, in accordance with the alphabet and, from another, paying attention to the inflected forms by means of affixation. Keeping in mind that there is a big difference in meaning between verbal forms formed by means of different affixes and the appropriate verbal noun, it is awkward to appropriately decide the alphabetical order of verbs. For example, if we consider the third person singular in the present or future indicative of verbs, the following forms will arise (15):

- (15) *mε-ur-s* 'goes', *muk'-ur-s* 'passes by', *ε-ur-s* 'descends vertically straight', *al-ur-s* 'descends onto something from the side, slantwise', *el-ur-s* 'ascends from the side of something', *ef-ur-s* 'ascends/rises', *mif-ur-s* 'enters the open space in the center', *min-ur-s* 'enters the enclosed space', *gim-ur-s* 'descends' etc.

and the quantity of dictionary entries with affixes attached to the same stem and reflecting different meanings will be more than enough. Otherwise, if we consider the appropriate verbal noun (16):

- (16) *-ula* 'going, walking'

to be the headword of the dictionary entry, different meanings of verbal constituents like those mentioned above will not be represented at the appropriate level.

To summarize, the normalization problems like alphabetization and lemmatization in the case of Megrelian are associated with the position of the verbal root in the verbal template.

3. Macro- and Micro-Structures of the Megrelian-English Dictionary

Any kind of electronic dictionary can be described as a database created with purpose to store and provide access to words or multi-word expressions (MWEs) including their meanings and translations. Generally, there are four major prerequisites (Atkins & Rundell, 2008; Gibbon & Van Eynde, 2000, and others) to the design of any dictionary: a) Linguistic specification (of the macrostructure and the microstructure); b) Database management system (DBMS) specification; c) Specification of the phases of lexicographic database construction: input, verification and modification; d) Presentation of and access to lexical information: access, re-formatting, dissemination. In the case of low-resourced endangered languages, three units are connected to the existence of special DBMS, which will allow not only to represent a dictionary but also to document all data including texts aligned with audio/video files and to process them. This system is to be used for the representation of linguistic specification, conversion of data to the appropriate format, its verification and modification and after all, the collected data should be accessed and disseminated online.

3.1 Macro-structure of the Megrelian-English dictionary

Taking into account that Megrelian-English dictionary is a result of a language documentation and corpus creation project, the Fieldworks Language Explorer

(FLeX, 2024) was used as DBMS system. FLeX is generally used in the context of lexicography and computational morphology due to its flexibility in configuring corpora, sketch grammars and dictionaries. So, FLeX was considered as a tool to suit the specific needs of the Megrelian language and its structure and to create the initial macro- and micro-structures of the Megrelian-English dictionary. FLeX not only enables the selection of the dictionary type but also allows the customization of the entry structure, including information on items that will be available in the output file for further processing and uploading to the portal.

For the bi-directional Megrelian-English dictionary, we determined the structure of dictionary entries, paying special attention to the output of FLeX. We revised the entries using a corpus-based approach, utilizing the annotated Megrelian corpus compiled in FLeX and consisting of 97,393 tokens (60,783 types). Additionally, we prepared a converter for the FLeX output to make it compatible with the lexical database using Python, converted the data to .sql file, and launched an online version of the dictionary.

3.2 Micro-Structure of the Megrelian-English Dictionary

The structures of dictionaries that could be generated from FLeX are the following: Hybrid forms, Lexeme-based, and Root-based. Given that we compiled the dictionary together with a corpus interface and paying attention to the problems mentioned above, especially lemmatization problems, special attention was paid to lexeme-based and root-based configurations.

- Lexeme-based: In a lexeme-based configuration, complex forms representing a single lexeme or a unit of meaning are used as the main entries. Lexemes are basic units of meaning, including their translations. This configuration is essential for languages with complex word structures, like Megrelian, where a single lexical unit encompasses various phonetic variants of grammatical forms and meanings. Thus, this configuration simplifies navigation by organizing entries based on core meanings (17).

(17) **ifɛnifɛn** (phon. var. **ifɛɛn; ifɛifɛn**) mod still

- Root-based: In a root-based configuration, root forms with complex forms as subentries are used as the main entries. Roots are considered the main morphemes, and in the case of Megrelian, they can consist of a single vowel or consonant depending on the part of speech. This approach may not be considered user-friendly, but it allows users to search for complex forms stemming from a common root and define the structure of separate words (18). Such an approach was used by Kipshidze in his printed dictionary (Kipshidze, 1914).

(18) **mifa-** (sp. var. **mfi-**; **nfa-**; phon. var. **mif-**; **mife-**; **mf-**; **mfa-**; **məf-**; **məfa-**; **məfe-**; **nf-**; **nfe-**) v:Preverb pfx PRV
mifartap cn turn
mifaular ptcp turn

Also, in the case of online dictionary, this approach allows searching not only for complex forms stemming from a common root, but also for separate inflectional morphemes and their meaning (19).

- (19) -იჲ (dial. var. -იჲჲ, phon. var. -ჲჲ) 1) n: Case sfx BEN 2) num: Case sfx BEN
3) adj: Case sfx BEN 4) pro-form: Case sfx BEN 5) ptcp: Case sfx BEN

The FLeX exporting function allows different options, especially, data can be represented as a) Configured Dictionary – Web page (XHTML); b) Dictionary, Reversal index – Webonary; c) Filtered Lexicon – LIFT 0.13 XML; d) Full Lexicon – LIFT 0.13 XML; e) Full Lexicon (lexeme-based) – SFM and f) Full Lexicon (root-based) – SFM. Taking into account that the lexeme- and root-based configurations were chosen for the online Megrelian-English dictionary, the special attention was paid to two formats:

- 1) Full Lexicon (lexeme-based) – SFM format, which allows exporting of the dictionary using Dictionary Formatter (MDF) lexeme-based standard (20):

- (20) \lx xḡiras
\lx_xmf xḡiras
\sn 1
\ps_en Temporal
\ps_kat dḡois zmnizēda
\g_en often
\sn 2
\ps_en Temporal
\ps_kat dḡois zmnizēda
\g_en frequently

- 2) Full Lexicon (root-based) – SFM, which provides exporting of the full lexicon using the Multi-Dictionary Formatter (MDF) root-based standard. In this format, subentries are included as part of the main entry rather than as separate entries with links to them (21).

- (21) \lx nat̃svl
\lx_xmf nat̃svl
\sn 1
\ps_en Main verb
\ps_kat mtavari zmna
\g_en replace

Both formats are compatible with Lexique Pro for publishing dictionaries online or in print. Additionally, both formats can be easily transformed into .sql format, which is important for integrating the dictionary into the portal. The transformation was made by a Python script specially developed for these purposes. A dictionary database (.db) file is linked to a FLeX corpus and its online version also is connected to the annotated Megrelian corpus interface. The dictionary database includes several key units of information, including the following core units:

- **Lexeme or Root Form:** The lexeme or root form represents the basic, uninflected or unmodified form of a word. It is the root or the lexeme form that serves as a reference point for various inflections, derivations, or variations of the word.

This form is, also, used for alphabetization purposes and allows users to look for the entries by pressing alphabet letters as well as to use the search option to look through the entries in alphabetical order.

- IPA (International Phonetic Alphabet): The IPA unit includes the phonetic transcription of the lexeme, allowing users to accurately pronounce the word.
- Gloss: This unit provides a brief, user-friendly explanation or translation of the meaning of a lexeme or a root into English. It serves as a quick reference in English for users to understand the sense of the word.
- Grammatical Information (Part of Speech): This unit specifies the grammatical category or part of speech to which the lexeme or the root belongs. It helps the user to understand the word's syntactic function and to look through its grammatical behaviour.
- Sense: The sense unit provides a detailed explanation of the different meanings or senses associated with a lexeme.

The connection between the dictionary and the corpus enriches the lexical entries with word usage context. The corpus interface allows users to explore how words are used in contextual variations across different genres and registers and, to determine their usage frequency. By analysing the occurrences of a word within the corpus, the users identify common and less common usages, especially in case of code-switches. This information helps researchers to identify the existence and importance of a word in everyday life. To summarize, at this moment the dictionary is available online at <https://xmf.iliauni.edu.ge/> and the connection between the dictionary and the corpus interface contributes to deep analysis of language patterns allowing users not only to see the meaning of words but also their morphosyntactic features.

3.3 Potential Users and Language Preservation Issues

Taking into account that the main scope of the project was to collect contemporary Megrelian data and to create an annotated corpus of Megrelian language with an online Megrelian-English dictionary, the potential users of the whole resource are diverse and include Megrelian speakers, who want to preserve their language and use vocabulary, language learners, who need a reliable resource for learning and translating Megrelian, linguists, researchers and students, who learn the grammatical structure of Megrelian and its contemporary situation. Also, the preservation issues can be ensured by regular updates to the corpus and the dictionary, compiling the comprehensive entries and maintaining traditional linguistic elements.

4. Conclusions

In conclusion, this lexicographic effort directed to the maintenance of low-resourced endangered Megrelian language is crucial for its preservation. Through the documentation of Megrelian language, lexicography becomes a keeper of linguistic heritage, collecting, categorizing, and preserving valuable linguistic data to prevent

its loss. The urgency of preserving Megrelian is provoked by the declining number of proficient speakers with each passing generation.

Also, the development of Megrelian-English dictionary plays an important role not only in the revitalization of Megrelian but also in the globalisation of project results. This dictionary freely accessible online can be considered as an important learning resource for those interested in its learning not only on the territory of Georgia, but worldwide.

The linking of the dictionary to the annotated corpus of Megrelian allows users not only to look through the dictionary entries but also to read different contexts expressing cultural knowledge, traditions, and identity. Such approach contributes to the cultural preservation of Megrelian as well. While from a linguistic perspective, this recourse contributes to the learning of language structures, grammar, and vocabulary.

References

- Atkins, S., & Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford University Press.
- Charaia, P. (1997). *Megrelian-Georgian Dictionary*. SPB.
- Chikobava, A. (2008 [1952]). *enat'mec'nierebis šesavali (Introduction to linguistics)*. t'bilis saxeimcip'o universiteti (Tbilisi State University).
- Eliava, G. (1999). *Megrelian-Georgian Dictionary*. Intellect.
- Gibbon, D., & Van Eynde, F. (2000). *Lexicon Development for Speech and Language Processing*. Kluwer Academic Publishers.
- Gippert, J. (1992). The Caucasian language material in Evliya Çelebi's 'Travel Book'. A Revision. In G. Hewitt (Ed.), *Caucasian Perspectives* (pp. 8–62). Lincom.
- Gippert, J. (2016). Complex Morphology and Its Impact on Lexicology: The Kartvelian Case. In T. Margalitadze, & G. Meladze (Eds.), *The XVII EURALEX International Congress* (pp. 16–37). Ivane Javakhishvili Tbilisi University Press.
- Gippert, J., Meurer, P., & Tandashvili, M. (2011–2024). *The Georgian national corpus*. Retrieved from The Georgian National Corpus: <http://gnc.gov.ge/>
- Glonti, A. (1964). *k'art'uli enis lek'sikologia (Georgian lexicology)*. c'odna (Knowledge).
- Glonti, A. (1983). *k'art'uli lek'sikograp'iis sakit'xebi (Questions in Georgian lexicography)*. sabčot'a sak'art'velo (Soviet Georgia).
- Kajaia, O. (2000–2009). *Megrelian-Georgian Dictionary, 4 volumes*. Nekeri.
- Kipshidze, I. (1914). *Megrelian-Russian Dictionary*. Typography of the Imperial Academy of Sciences.

- Klimov, G., & Kajaia, O. (2023). *Megrelian-Russian-Georgian Dictionary*. Govorun.
- Kobalia, A. (2010–2020). *Megrelian Dictionary*. Artanuji.
- Lobzhanidze, I. (2019). Computational Model of the Modern Georgian Language and Search Patterns for an Online Dictionary of Idioms. *Language, Logic, and Computation. TbiLLC 2018. Lecture Notes in Computer Science*, vol 11456, 187–208.
- Pochkhua, B. (1974). *k'art'uli enis lek'sikologia (Georgian Lexicology)*. t'bilisis saxelmcip'o universiteti (Tbilisi State University).
- Sarjveladze, Z. (1997). *zveli k'art'uli ena (Old Georgian Language)*. Tbilisi State Pedagogical University Press.
- Shanidze, A. (1976). *zveli k'art'uli ena (Old Georgian Language)*. t'bilisis saxelmcip'o universiteti (Tbilisi State University).
- SIL Feildworks Language Explorer (FLeX), Version 9.1. Retrieved January 1, 2024, from <https://software.sil.org/fieldworks/>
- Standard ISO/DIS 24611 (2012). *Language Resource Management – Morpho-syntactic Annotation Framework (MAF)*, No 24611. <https://www.iso.org/standard/51934.html>
- UNESCO: Moseley, Ch. & Nicolas, A. (2021). *World Atlas of Languages: Mingrelian*. Retrieved from <https://en.wal.unesco.org/languages/mingrelian>

Acknowledgements

This work was supported by Shota Rustaveli National Science Foundation of Georgia (SRNSFG) [grant number: FR-21-993-3, 2021-2025]. The authors would like to express their gratitude to the two anonymous reviewers for the many helpful comments and suggestions and to the organizing committee of the EURALEX for their support.

Contact information

Irina Lobzhanidze

Ilia State University, Institute of Linguistic Studies
 irina_lobzhanidze@iliauni.edu.ge

Rusudan Gersamia

Ilia State University, Institute of Linguistic Studies
 rgersamia@iliauni.edu.ge

Nino Tsulaia

Ilia State University, Institute of Linguistic Studies
 nino.tsulaia.1@iliauni.edu.ge

Chenlu Yu

A GERMAN-CHINESE E-DICTIONARY OF MANUFACTURING TECHNOLOGY IN AUTOMOTIVE INDUSTRY

Entry Design

Abstract We intend to create an online dictionary for manufacturing technology in the automotive industry, which will be available in German and Chinese. The dictionary is designed to improve specialized communication, and its target users include Chinese engineers, students, and interpreters in this sector. The development of entry structures using Frame-Based Terminology (FBT) as a dynamic process-oriented approach (cf. Faber et al., 2005; Faber, 2011) will be the main topic of this paper. Specifically, we will present a corpus-based methodology for integrating contextual information, especially syntactic-semantic features, into the entries. At this point, we start with the description of predicative terms. It might be conceivable to provide a semantic-based approach that can serve as a practical solution for future bilingual dictionaries in various engineering disciplines.

Keywords specialized lexicography; bilingual dictionaries; frame-based terminology; dictionary entries

1. Introduction

In recent decades, German original equipment manufacturers and automotive suppliers have set up overseas offices and production bases in China. Statistics show that Germany is rapidly becoming a top choice for Chinese students majoring in technology and engineering. Furthermore, as China and Germany collaborate in various of industrial sectors, skilled translators with in-depth knowledge are becoming increasingly important.

In this context, a high-quality German-Chinese automotive manufacturing dictionary is extremely useful. In contrast to the traditional approach used in most bilingual specialized dictionaries for the Chinese market, which typically focuses on collecting nouns and offers a static micro-structure, our dictionary incorporates exclusive predicative terms, like specialized verbs with their participants and their contextual information, into the dictionary entries. This approach recognizes the need for effective daily communication in the automotive manufacturing industry, especially when it comes to understanding and refining German technical documentation and manuals and facilitating the acquisition of subject-specific knowledge. According to ten Hacken (2016, p. 73), there are three key questions to consider when applying theories to word meanings in bilingual dictionaries:

- a) How can a dictionary entry be interpreted?
- b) How can we determine how good a dictionary entry is?
- c) How can we compile a dictionary entry?

These three questions are fundamental to our research project. In this paper, we introduce the current lexicographic stage of the project and offer a preliminary prototype of an entry. Section 2 summarizes the project, including the target user group, the dictionary's intended function, the constitution of the corpora, and the types of headwords included. Section 3 outlines the methodology for compiling a sample entry, including constructing definition patterns, collecting contextual information, and identifying equivalents within the entry. Section 4 presents a sample entry based on the outcomes of the preceding methodological steps. Finally, Section 5 concludes with a summary and offers recommendations for future work.

2. The Outline of the Project

2.1 The User Group

Schierholz (2003; 2014) and Gouws (2020) argue that specialized dictionaries are designed as reference tools for experts, laypeople, or learners in specific fields. By including specialized words within a certain subject field, these dictionaries assist in bridging language gaps and prompting effective communication. Given the widespread accessibility of online dictionaries, the target users include three groups: (1) Chinese engineers in the automotive and related industries; (2) students majoring in automotive engineering, mechanical engineering, and related subjects; and (3) translators working in the mechanical, automotive, and other technical sectors.

2.2 Functions of the Dictionary

Most specialized dictionaries and terminology databases prefer to emphasize definitions and concept relationships rather than starting from the user's perspective. According to Bergenholtz & Tarp (2010), the primary goal of any lexicographic or terminological product should be to address the users' needs. Considering the diverse characteristics of users, our dictionary should be multi-functional, accommodating the needs of three user groups involved in text comprehension, production, and translation. For instance, engineers frequently encounter specialized communicative contexts in which they must interpret German technical drawings and documents and communicate with their German counterparts via email. In such scenarios, having access to equivalent terms and contextual information is essential for accurate comprehension and effective communication.

2.3 Corpora

We are developing two comparable corpora consisting of handbooks and manuals pertinent to the automotive sector, academic literature and books used as teaching materials, and papers written by professors and experts.

Currently, three *Handbooks for Automotive Management* published by International Automotive Groups, along with three academic books, are included in each of the German and Chinese corpora. In the future, we plan to add papers and articles to each corpus covering topics related to automotive manufacturing, such as manufacturing technology, electric automobile production, vehicle repair, etc.

To sum up, the German corpus contains 750,867 tokens, equivalent to 565,143 words, while the Chinese corpus consists of 883,141 tokens, corresponding to 734,127 words. Our goal is to reach the size of 1 million words for each corpus.

2.4 The Type of Headwords

Scholars have recently recognized that employing predicative units to represent processes, events, and attributes can improve the utility of specialized dictionaries and terminological resources. Unlike nouns that refer to entities, verbs can be analyzed in terms of their argument structures and the semantic roles of each participant. This approach allows for the construction of a conceptual paradigm that is based on both syntax and semantics (cf. L'Homme, 2012; 2015), and it has been proven that concentrating on specialized verbs can be broadly implemented in specialized dictionary projects and knowledge bases, such as those in the medical field (cf. Wandji et al., 2013) and the legal domain (cf. Pimentel, 2012). Specialized verbs, morphological nominalizations and compounds will be considered for headwords inclusion in our project. In addition to single terms, many multi-word terms paired with certain verbs through concordance and clustering could also be added into headwords.

3. Methodology

We employ a bottom-up approach, starting with the selection of specialized verbs from corpora (Section 3.1). Then we define frame categories and definition patterns to describe specialized verbs (Section 3.2), annotate syntactic-semantic information using concordance analysis (Section 3.3), and finally assign the equivalent (Section 3.4). The data collection and pre-processing are completed with Sketch Engine.

3.1 Term Selection

Before term selection, we divide the term candidates into three stages of the automotive production process: pre-production, in-production, and post-production. Equipment installation and setup, as well as material preparation, are all part of the pre-production stage, which can also occur throughout production. Therefore, sub-categories such as *manipulation* and *preparation* are essential at these stages. According to the German Industry Standard DIN 8580 (cf. Förster & Förster, 2018), the manufacturing process within the production stage is organized into six sub-categories: primary shaping, material forming, dividing, joining,

modifying material property, and coating. The post-production stage includes sub-categories such as testing and shipment, among others.

Using frequency indicators suggested by L'Homme (2015), we examine the frequency of verb occurrences in their context based on this classification to find potential term candidates. To be specific, we first go through verbs that have an absolute frequency value greater than 100 and work with industry experts to perform the preliminary filtering, producing a list of verbs for each language. Following that, each of these potential verbs is examined separately to find multi-word term candidates using clustering and concordance. This process is used for both corpora in the same way.

After completing these procedures, we chose the German predicative term *einstellen* 'to regulate' as a candidate term. As seen in the corpora, this candidate term has a frequency value of 127 in the German corpus and a frequency value of 313 in the Chinese corpus.

3.2 Frame-Based Definition Pattern

Frame Semantics (cf. Fillmore, 1982; 1985) proposes that to understand the meaning of a word, one must activate the associated frame and the comprehensive knowledge involved in the frame. Expanding upon Fillmore's notion, Faber (2011) defined frames in a broader sense to describe knowledge structures. In addition to enriching the dynamism and coherence of the terminological knowledge base, this active, process-oriented approach extends the conceptual description beyond the hierarchical structure.

Based on concordance analysis (see Figure 1), the participants of the predicative term *einstellen* 'to regulate' are classified into two types:

- (1) Concrete tools or devices: the spring, machine, gear, and valve can be adjusted, as shown in items 2, 5, 7, and 9.
- (2) Abstract units, like parameters: items 1, 3, 4, 8, 10, and 14 indicate that the frequency, limit, and time can be regulated.

A human being performs either (1) or (2) of these tasks at the initial production stage or while the machine is operating. As a result, we designate the sub-category *manipulation* as the frame for characterizing the predicative term *einstellen*. Three types of conceptual relations are included in the definitional pattern for the frame *manipulation*: (1) *is_a*; (2) *has_patient*; and (3) *has_instrument*. Within this frame, other related actions at the pre-production stage, such as *festlegen*, *überprüfen*, and *kalibrieren*, share similar features and argument templates.

	Lemma	Frequency	Relative [?]
1	☐ Grenze einstellen	11	14.65 ***
2	☐ Ventil einstellen	3	4.00 ***
3	☐ Betriebsgrenze einstellen	3	4.00 ***
4	☐ Drehzahl einstellen	2	2.66 ***
5	☐ Maschine einstellen	2	2.66 ***
6	☐ Mann einstellen	2	2.66 ***
7	☐ Wechselzahnrad einstellen	2	2.66 ***
8	☐ Werkstückhöhe einstellen	2	2.66 ***
9	☐ Anpressfeder einstellen	1	1.33 ***
10	☐ Zeit einstellen	1	1.33 ***
11	☐ Viskosität einstellen	1	1.33 ***
12	☐ Drahtvorschub einstellen	1	1.33 ***
13	☐ Gegenkoppelnetzwerk einstellen	1	1.33 ***
14	☐ Stromfrequenz einstellen	1	1.33 ***
15	☐ Iuftstromes einstellen	1	1.33 ***
16	☐ Schweißstromstärke einstellen	1	1.33 ***
17	☐ Stückzahl einstellen	1	1.33 ***

Fig. 1: The participants of the predicative term *einstellen*

3.3 Syntactic-Semantic Contextual Information

Contextual information plays a crucial role in language comprehension and production. Domínguez Vázquez & Gouws (2023) highlight the importance of contextual information, noting that speakers combine words with contextual cues in various communicative situations. In the Frame-based approach, contextual information can be categorized by its scope (local and global) and the type of information conveyed (syntactic, semantic, and pragmatic variables) (cf. Faber et al., 2016). We focus on extracting local contextual data and generating syntactic-semantic patterns using CQL queries in Sketch Engine. This process involves the following steps:

- (1) **Select predicative units:** Currently, we focus on specialized verbs.
- (2) **Define participants and the valence structure:** Identify the participants and the valence structure syntactically linked to the verb, including prepositional phrases, to generate prototypical patterns.
- (3) **Extract data using CQL Queries:** Employ CQL queries in Sketch Engine to gather data for each prototypical pattern and identify the relevant slots.
- (4) **Annotate part of speech and morphological variation:** Annotate the part of speech for the extracted data and document any morphological variations of the verb.
- (5) **Establish semantic roles and add examples:** Define the semantic roles for each valence slot and include example sentences.

3.4 Equivalents Assignment

After grouping the German verbs into frames and completing their descriptions, we plan to link these verbs to their Chinese equivalents using the frame labels. The criteria for selecting equivalents as outlined by Pimentel (2012, p. 1793) include (1) the way the verbs evoke the frame; (2) the actantial structures of the verbs; and (3) the syntactic patterns of the verbs.

4. A Sample Entry

Let’s address each of the three questions mentioned in Section 1:

- a) Our dictionary entries contain at least eight item categories: headwords, grammatical information, equivalents, frames, definitions, semantic roles and relations, syntactic-semantic structures, and examples extracted from corpora (see Figure 2).
- b) A comprehensive dictionary entry must address users’ needs. When encountering technical texts in German, Chinese equivalents should facilitate comprehension promptly. The provided syntactic-semantic patterns aid users in completing writing tasks efficiently. Additionally, the description of verbs, especially in the highlighted section (yellow and grey sections) of the entry, serves as a concise knowledge structure to help users understand and recognize concepts.
- c) The Frame-based approach, which integrates semantic and syntactic behaviors, has proven to be an effective method for developing bilingual specialized dictionary entries, especially for predicative terms such as verbs.

einstellen (Eng: to regulate) vt.

Definition 定义: etwas mit Hilfe von etwas zu regeln

Chinese equivalent 中文: 调节, 设置

[IS_A] Handlung

Frame Category 类别: Manipulation

Semantic network 语义组合

[Object] Zubehörteil, Parameter

[Instrument] Ersatzteil, Tool

Structure and example 结构和例子

Object [Gerät, Zubehörteil, Parameter] (in Akkusativ) + verb

	German	Chinese
Example	Grenze; Frequenz	范围; 频率
	Maschinen; Zylinder	机器; 气缸

Preposition + Object [Werkzeug, Ersatzteil] (in Dativ) + verb

	German	Chinese
Example	mit Ventil	阀

Sentences 例句

1. Damit das Motormanagement ein definiertes Luft-Kraftstoff-Gemisch einstellen kann, muss diese Luftmenge genau bekannt sein. 必须精确地知道这个空气量, 以便发动机管理系统能够调节一个确定的空气/燃油混合气量.
2. Bei der Steuerung mit Pro-portionalventil werden alle Geschwindigkeiten mit diesem Ventil eingestellt. 在带比例阀的控制系统中, 所有速度均由该阀设置.

Fig. 2: The Sample entry for the verb *einstellen*

5. Conclusion and Future Work

Compiling a bilingual dictionary for scientific and technical terms is not a new topic in specialized lexicography. Our contribution is to use a Frame-based approach to improve the dynamism and coherence of entries in German-Chinese bilingual dictionaries in the field of the automotive industry.

Regarding the limitations, the next step involves broadening experts' articles and papers into the comparable corpora and further refining the microstructure of the dictionary entries. The future dictionary will contain 50 preliminary predicative terms and the comprehensive entries as shown in Figure 2. A number of multi-word terms will be taken into account in headwords as well.

References

- Bergenholtz, H., & Tarp, S. (2010). CHAPTER 1. LSP Lexicography or Terminography? The lexicographer's point of view. In P. A. Fuentès-Olivera (Ed.), *Specialised Dictionaries for Learners* (pp. 27–38). De Gruyter.
- Domínguez Vázquez, M. J., & Gouws, R. H. (2023). The Definition, Presentation and Automatic Generation of Contextual Data in Lexicography. *International Journal of Lexicography*, 36(3), 233–259. <https://doi.org/10.1093/ijl/ecac020>
- Faber, P., Márquez Linares, C., & Vega Expósito, M. (2005). Framing Terminology: A Process-Oriented Approach. *Meta*, 50(4). <https://doi.org/10.7202/019916ar>
- Faber, P. (2011). The dynamics of specialized knowledge representation: Simulational reconstruction or the perception–action interface. *Terminology*, 17(1), 9–29.
- Faber, P., Reimerink, A., & Leon-Araúz, P. (2016). Encoding context in bilingual specialized dictionaries. In T. Margalitadze, & G. Meladze (Eds.), *Lexicography and Linguistic Diversity: Proceedings of the XVII EURALEX International Congress* (pp. 491–498). Ivane Javakhishvili Tbilisi State University.
- Fillmore, C. J. (1982). Frame Semantics. In *Linguistics in the Morning Calm: Selected Papers from SICOL-1981* (pp. 111–137). Hanshin Publishing Co.
- Fillmore, C. J. (1985): Frames and the Semantics of Understanding. *Quaderni di Semantica*, 6, 222–253.
- Förster, R., & Förster, A. (2018). Einteilung der Fertigungsverfahren nach DIN 8580. In R. Förster, & A. Förster (Eds.), *Einführung in die Fertigungstechnik* (pp. 23–135). Springer. https://doi.org/10.1007/978-3-662-54702-1_2
- Gouws, R. H. (2020). Special Field and Subject Field Lexicography Contributing to Lexicography. *Lexikos*, 30(1), 143–170. <https://doi.org/10.5788/30-1-1568>
- L'Homme, M.-C. (2012). Adding syntactico-semantic information to specialized dictionaries: an application of the FrameNet methodology. *Lexicographica*, 28, 233–252. <https://doi.org/10.1515/lexi.2012-0012>

L'Homme, M.-C. (2015). Predicative Lexical Units in Terminology. In N. Gala, R. Rapp, & G. Bel-Enguix, (Eds.), *Language Production, Cognition, and the Lexicon. Text, Speech and Language Technology*, vol. 48 (pp. 75–93). Springer, Cham.

Pimentel, J. (2012). Identifying equivalents of specialized verbs in a bilingual comparable corpus of judgments: A frame-based methodology. In N. Calzolari et al. (Eds.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)* (pp. 1791–1798). European Language Resources Association (ELRA).

Schierholz, S. J. (2003). Fachlexikographie und Terminographie. *Zeitschrift für Angewandte Linguistik*, 39, 5–28.

Schierholz, S. J. (2014). 17. New developments in lexicography for special purposes I: An overview of linguistic dictionaries. In R. Gouws, U. Heid, W. Schweickard, & H. Wiegand (Eds.), *Supplementary Volume Dictionaries. An International Encyclopedia of Lexicography: Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography* (pp. 431–442). De Gruyter Mouton.

ten Hacken, P. (2016). Bilingual Dictionaries and Theories of Word Meaning. In T. Margalitadze, & G. Meladze (Eds.), *Lexicography and Linguistic Diversity: Proceedings of the XVII EURALEX International Congress* (pp. 61–76). Ivane Javakhishvili Tbilisi State University.

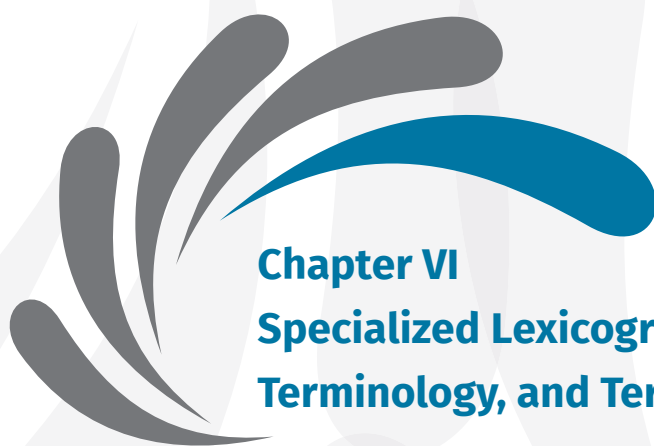
Wandji, O., L'Homme, M.-C., & Grabar, N. (2013). Discovering Semantic Frames for a Contrastive Study of Verbs in Medical Corpora. In *Actes de la 9e conférence internationale Terminology and Artificial Intelligence (TIA'13)*. Paris, France.

Contact information

Chenlu Yu

Universitat Pompeu Fabra

chenlu.yu@upf.edu



Chapter VI
Specialized Lexicography,
Terminology, and Terminography

Andrea Abel and Natascia Ralli

GENDER IN ELECTRONIC DICTIONARIES AND TERMINOLOGY DATABASES

State of the Art and Future Directions

Abstract What strategies are currently being applied in electronic dictionaries and terminology databases to gender representation, with a particular focus on feminine agentives? Starting with an overview of the state of the art as to gender studies in lexicography and terminology science, in this paper we use a case study to reflect upon the approaches in electronic dictionaries and terminology databases, collecting both good practices and concrete limitations that hinder an equal gender representation by contrasting different linguistic and socio-cultural contexts. Since lexical resources can be considered not only utility instruments but also “agents” that reflect society (Gouws, 2022, p. 40) there is a risk that they contribute, for example, to perpetuate gender stereotypes (Müller-Spitzer, 2023, p. 80). Thus, it is important for lexicographers and terminologists to discuss and establish methods for avoiding gender bias in lexical resources.

Keywords lexicography; terminology work; gender

1. Introduction

With the rise of the feminist movement in the 1960s, the attention towards all forms of sexism, including linguistic sexism, became an object of scientific and popular discussions. 60 years later, the debate has not dried up yet. On the contrary, in recent years, it has regained the general public’s interest, especially through the #MeToo movement (Cavagnoli & Dragotto, 2021, p. 1) and an increasing devotion of the media to gender inequalities in general. Also, regarding language, a steadily growing number of questions have been posed by the public to academies’ experts (e.g., Accademia della Crusca; RAE@informa).

In the field of lexicography, since the 1970s, numerous authors analyzed and reflected upon gender treatment in dictionaries (e.g., Graham, 1975; Hampares, 1976; Pusch, 1984; Whitcut, 1984; Calvo Ramos, 1998; Nübling, 2009; Westveer et al., 2018; Fuertes-Olivera & Tarp, 2022; Müller-Spitzer & Lobin, 2022). Discussions in this field have led some dictionaries to revise their approach, deciding to make the feminine agentives, i.e., “linguistic forms that indicate an agent, as job titles, etc.” (Bengoechea, 2017, p. 200),¹ visible as separate entry or to record the masculine and feminine agentives on the same level (e.g., *Duden*, *Diccionario de la lengua española* of the Real Academia Española, *Oxford English Dictionary*, *Treccani*).

¹ The concept “agentive” also includes professions, jobs, roles, and educational qualifications.

In terminology work, the debate on how feminine agentives should be recorded is still open (Ralli & Evers, in review). The problem is methodological at base: concept is the core of terminology science and, as such, is not related to gender. However, there is the need to record feminine agentives in terminology databases to be able to use them correctly and consistently also in combination with, for example, proofreading tools, machine translation engines or computer-assisted translation (CAT) tools (Evers, 2022, p. 13).

Against this background, we want to address the following question in this paper: What strategies are currently being applied in electronic dictionaries and terminology databases to gender representation, with a particular focus on feminine agentives? We address this question by describing concrete examples from mainly two different socio-cultural contexts, the German-speaking area and the Italian-speaking area. We concentrate on the German dictionary *Duden online* and the Italian dictionary *Treccani* (online version) for the exemplary analyses of electronic dictionaries. The terminology databases, on the other hand, are *LinguaPC*, *TERMDAT*, and *TERMCAT*. Through this analysis, we pursue two main objectives. Firstly, we aim to summarize the state of the art. Secondly, we describe, by means of exemplary illustrations, the current strategies applied in electronic dictionaries and terminology databases to represent feminine agentives.² We concentrate on both their macrostructure (lemmatization) and their microstructure (definitions and examples), taking into account the different methodological approaches, a predominantly semasiological one in lexicography versus a predominantly onomasiological one in terminology work which leads to different foci.

2. State of the Art in Lexicography

2.1 German-Speaking Area

The well-known article “Sie sah zu ihm auf wie zu einem Gott – Das DUDEN-Bedeutungswörterbuch als Trivialroman”³ by Pusch (1984) is usually cited as the starting point for the discussion of gender issues in German lexicography. In it, Pusch examined part of the first edition of the dictionary from 1970 and found, for example, that only a small proportion of the personal names refer to women and that the role of women is strongly orientated towards men and children. Women are often portrayed either as virtuous and naive or as seductive and cunning (see Kotthoff & Nübling, 2018, pp. 180–181). In subsequent editions, the Duden editors responded – to varying degrees – to the criticism. The Duden editors themselves described in an article (Kunkel-Razum, 2012, pp. 2015–2016), in the title of which they referred directly to Pusch with “Er sah zu ihr auf wie zu einer Göttin”,⁴ how the theoretical discussion in the context of gender studies was incorporated into practical dictionary work. In 2002, for example, female personal names were included in all entries for

² We have limited the analysis to binary forms.

³ “She looked up to him like a god – The DUDEN dictionary of meanings as a light novel” (translation by the authors).

⁴ “He looked up to her like to a goddess” (translation by the authors).

male personal names – with *Päpstin*⁵ as the only exception (however, included in 2009⁶) – to make women more visible in dictionaries. More revisions followed in response to criticism by Porsch (2004), for example, in an article also alluding to Pusch, as “Das DUDEN-Universalwörterbuch als Fortsetzung eines Trivialromans”⁷ reveals in the title.

A comparative overview since the 1970s by Kotthoff & Nübling (2018, pp. 180–184) shows that stereotypical role models can also be found in German dictionaries from other publishers, with only the *Wörterbuch der deutschen Gegenwartssprache* (1961–1977) from the former GDR partially breaking through the usual stereotypes. More recent online dictionaries also reproduce traditional role models, as Thüne & Leonardi (2006) show using the *German Word Thesaurus*.

The fact that dictionaries reflect stereotypes, i.e., human thinking in group categories, also with regard to gender, cannot initially be blamed on them. After all, stereotypes facilitate information processing (Müller-Spitzer & Lobin, 2022, p. 36). On the other hand, even if dictionaries are intended to be descriptive, they are often perceived as normative authorities. Accordingly, lexicographers have a great responsibility in the design of lexicographical entries, which inevitably convey a certain worldview (Müller-Spitzer & Lobin, 2022, p. 37; Kunkel-Razum, 2012, p. 216). Based on a case study with German newspaper corpora, Müller-Spitzer & Lobin (2022) show a comparatively strong presence of gender stereotypes in this genre, which in turn affect the lexicographical products, e.g., *Duden online* or *lexiko*, especially their automatically generated parts (e.g., typical word combinations in *Duden online*). *Duden online* has been available since 2011. We use it as an example of good practice as it changed and expanded its offering in 2021 with regard to feminine personal and occupational words.⁸

2.2. Italian-Speaking Area

Several authors have studied and analyzed gender biases in dictionaries over the last decades.⁹ At the end of the 1980s, the very first Italian contribution to sexism in language was authored by Alma Sabatini (1987). She (1987, p. 28–29) detected numerous asymmetries between meanings and a strongly stereotyped depiction of women.

Lepschy et al. (2001) believed that progressive choices in lexicography are not always suitable. A term might be considered insulting because dictionaries say so rather than because it is really used as insulting (Lepschy et al., 2001, p. 8).

⁵ Eng. ‘Popess’.

⁶ Unpublished talk from Kathrin Kunkel-Razum, editor-in-chief of the Duden publishing house on “Arzt, der – männliche Person, die ...; Ärztin, die – weibliche Person, die ... Women and men in current German-language meaning dictionaries”, EAAL Conference on 16–17 June 2022, Tallinn.

⁷ “The DUDEN explanatory dictionary as a sequel to a light novel” (translation by the authors).

⁸ Unpublished talk from Kathrin Kunkel-Razum 2022 (see above).

⁹ For a broader overview, please refer to Fusco (2012).

Fusco (2009) shared Sabatini's position on dictionaries' role in society. Her analysis of more than 1,300 lemmas of the *GRADIT* dictionary related to the word *donna*¹⁰ revealed that in many occurrences, women are depicted through stereotypes, especially as to physical and moral attributes as well as to out-of-date traditions. The intellectual value of women was heavily misrepresented (Fusco, 2009, pp. 212–217). In 2019, she analyzed the *GRADIT* dictionary again, finding out that 30 years after the publication of Sabatini's recommendation, many profession names were still not accepted even if documented in dictionaries (e.g., *architetta*, *poeta*, *ministra*¹¹). Women were reluctant to employ professional denominations because they did not feel representative of themselves. They still tended to prefer the masculine noun, as it better represented the function they finally came to occupy, according to Fusco (2019, p. 44–46), as well as to Sabatini (1987, p. 30).

Furthermore, Manera & Bazzanella (2006) delved into synonyms, antonyms, and definitions of the *Word Thesaurus* for a selection of terms. Their analysis revealed an evident asymmetry in masculine and feminine representation, severely biased towards women. For instance, synonyms of the word *donna* referred to Christian-practice terms, physical (as a sexual object), moral attitudes, or specific roles (wife, mother). Furthermore, “with regard to the number of meanings, synonyms, and antonyms, the quantitative asymmetry is markedly on the side of the woman” (*Ibid.*, p. 117).

Finally, it is worth mentioning a new feature in the new edition of Treccani's *Dizionario della lingua italiana*, published in 2022, which the publisher describes as “revolutionary” in its advertising of the print version.¹² This includes the lemmatization of feminine forms of nouns and adjectives, which were traditionally only recorded in the masculine form. We will include the online version of this dictionary as a second example of good practice in our study, even if it is not clear from the information available to us how many changes have also found their way into the online edition.

3. E-dictionaries Today

In this section, we will describe by means of two case studies, the current strategies adopted to represent the feminine world in online dictionaries. We will start by analyzing examples of the recent reform that occurred in the *Online-Duden*. Then, we will present analogous examples from *Treccani*, an Italian dictionary breaking new ground in terms of feminine representation, both in its micro- and macrostructure.

Based on the current state of research, the following aspects are in particular considered: a) the presence of female agentives, b) lemmatization, c) descriptions of meaning and examples, and d) editorially vs. automatically created parts in dictionary entries.

¹⁰ Eng ‘woman’

¹¹ Feminine form of *architect*, *poet*, and *minister*.

¹² https://emporium.treccani.it/it/il-dizionario-italiano-3012242.html?utm_source=TreccaniIT&utm_medium=ban-ner&utm_campaign=Dizionario_dell_Italiano

When selecting the lemmas for German, we look for examples according to the following linguistic possibilities for gender specification: derivation (-*in* suffix), composition (-*frau*), and substantivization of participles (-*te*, -*nde*) (see Kotthoff & Nübling, 2018, p. 130; Westweer et al., 2018, p. 375). For Italian, on the other hand, we consider the following: derivation (-*a*, -*essa*, -*trice* as suffixes, zero suffix, i.e., identical masculine and feminine forms: -*ante*, -*ista*¹³) (see Fusco, 2019, pp. 36–37).

3.1 German

For the exemplary analysis, the following nouns were analyzed in *Duden online*: *Anwalt/Anwältin*, *Autor/Autorin*, *Architekt/Architektin*, *Arzt/Ärztin*, *Bäcker/Bäckerin*, *Friseur/Friseurin*, *Feuerwehrmann/Feuerwehrrfrau*, *Abgeordneter/Abgeordnete*, *Vorsitzender/Vorsitzende*, *Studierender/Studierende*.¹⁴

The descriptions of meaning were adapted for both genders in the course of the revision of *Duden online* by essentially replacing the description of “jemand, der [...]” with “männliche Person, die [...]” on the one hand and introducing “weibliche Person, die [...]”¹⁵ on the other, for example in the lexemes *Arzt* and *Ärztin*. Other examples are comparable, such as *Anwalt* or *Anwältin*, whereby in this case the detour via the explanation of the meaning by means of synonyms *Rechtsanwalt* or *Rechtsanwältin* is to be taken, which in turn is introduced with “Jurist [bzw. Juristin] mit staatlicher Zulassung [...]”.¹⁶ Finally, *Jurist* or *Juristin* is described according to the usual pattern of “männliche [or weibliche] Person, die [...]”. The same applies to *Architekt* or *Architektin*, who are first introduced as *Fachmann* or *Fachfrau*,¹⁷ while only the entries for the latter follow the familiar pattern. This approach reveals a strategy whereby only one definition of semantically related lemmas follows the usual pattern and tends to be the one in which the *genus proximum* is explained.

In the case of masculine agentives whose feminine form ends in -*in*, as well as substantivized participles, a separate box on the use of the word is inserted, in which the problem of the generic masculine is pointed out, but without using this technical term: “In bestimmten Situationen wird die maskuline Form (z. B. *Arzt*, *Mieter*, *Bäcker*) gebraucht, um damit Personen aller Geschlechter zu bezeichnen. Bei dieser Verwendung ist aber sprachlich nicht immer eindeutig, ob nur männliche Personen gemeint sind oder auch andere. Deswegen wird seit einiger Zeit über sprachliche Alternativen diskutiert”¹⁸ (emphasis in original). The use of the substantive present participle for agentives (see e.g., Müller-Spitzer, 2022), which is particularly

¹³ -*tora*, rare, and ignored

¹⁴ Masculine and feminine forms of *lawyer*, *author*, *architect*, *physician*, *baker*, *hairdresser*, *fireman*, *member of parliament*, *chairperson*, and *student* [literally ‘studying male/female person’]

¹⁵ Eng. ‘Someone who’, ‘male person who’, ‘female person who’.

¹⁶ Eng. ‘[Male and female] lawyer with a state licence’.

¹⁷ Male and female expert.

¹⁸ “In certain situations, the masculine form (e.g., *doctor*, *tenant*, *baker*) is used to refer to people of all genders. However, it is not always clear from a linguistic point of view whether only male persons are meant or others as well. For this reason, linguistic alternatives have been under discussion for some time” (translation by the authors); see e.g., <https://www.duden.de/rechtschreibung/Arzt>.

controversial in public and linguistic discourse, is provided with an additional box “Besonderer Hinweis”¹⁹, which appears in the article in the form section at the beginning as an announcement, and finally in the meaning section with the note itself: “Als geschlechtsneutrale Bezeichnung oder als Ausweichform für die Doppelnennung *Studentinnen* und *Studenten* setzt sich der Plural *Studierende* immer mehr durch”²⁰ (emphasis in original).

In the lexicographical examples, a quantitative difference between masculine vs feminine agentives is noticeable: In most cases, there are fewer examples for feminine forms. In the entries *Abgeordneter* and *Abgeordnete*, the number is identical, with only the personal designation changing in the examples themselves. In contrast, in the other entries analyzed, the examples are only identical to a small extent. The attribute *behandelnd*²¹, for example, describes both *Arzt* and *Ärztin*, while the attribute *leitender*²² only describes *Arzt*. However, such stereotypical role allocations appear to be rare overall. Variations in meaning, such as *Arzt* or *Ärztin* as an individual vs. as an institution, are not differentiated in the meaning descriptions. They have to be inferred from the examples, e.g., “zum Arzt gehen”²³, for which there is no female equivalent.

Major differences can be recognized in the automatically generated “Typische Verbindungen” for male vs. female agentives. Firstly, it can be seen that in some cases, there are no typical co-occurrences for the female terms at all, e.g., *Architektin* and *Bäckerin*. It is noticeable that more co-occurrences between male agentives and other male agentives are shown than with female ones; the opposite is true for female agentives. If we take another closer look at the example of *Arzt* or *Ärztin*, we can see that for *Arzt* there are links to *Apotheker*, *Psychologe*, *Anwalt*, *Patient* and *Ärztin*, whereas for *Ärztin* there are links to *Psychologin*, *Psychiaterin*, *Juristin*, *Rechtsanwältin*, *Pädagogin*, *Krankenschwester* and *Arzt*. There are also differences concerning the adjectives: With *Arzt* we find *jung*, *tätig*, *jüdisch*, *praktisch*, *neu*, *niedergelassen*, *diensthabend*, *tätig*²⁴, whereas with *Ärztin* we find *angehend*, *hübsch*, *blond*, *jung*, *praktisch*, *niedergelassen*, *diensthabend*.²⁵ The verb combinations are also different: *Arzt* has *verordnen*, *raten*, *rufen*, *verschreiben*, *sagen*, *aufsuchen*, *behandeln*, *diagnostizieren*,²⁶ while *Ärztin*, by contrast, has *arbeiten*, *empfehlen*, *fragen*, *sagen*, *meinen*, *antworten*, *verheiraten*, *verlieben*.²⁷ A number of particularly female gender stereotypes are depicted here, such as *hübsch*, *blond* and *heiraten*.²⁸

¹⁹ Eng. ‘Special note’; also for the feminine form.

²⁰ “As a gender-neutral term or as an alternative form for the double mention *female students* and *male students*, the plural “*studying people*” [literal translation] is becoming increasingly popular” (translation by the authors); see e.g., <https://www.duden.de/rechtschreibung/Studierende>.

²¹ Eng. ‘attending [physician]’.

²² Eng. ‘head [physician]’.

²³ Eng. ‘to go to the doctor’.

²⁴ Eng. ‘young’, ‘active’, ‘Jewish’, ‘practical’, ‘new’, ‘established’, ‘serving’, ‘active’.

²⁵ Eng. ‘trainee’, ‘pretty’, ‘blonde’, ‘young’, ‘practical’, ‘settled’, ‘serving’.

²⁶ Eng. ‘prescribe’, ‘advise’, ‘call’, ‘prescribe’, ‘say’, ‘visit’, ‘treat’, ‘diagnose’.

²⁷ Eng. ‘work’, ‘recommend’, ‘ask’, ‘say’, ‘mean’, ‘answer’, ‘marry’, ‘fall in love’.

²⁸ Eng. ‘pretty’, ‘blonde’, ‘marry’.

3.2 Italian

The following Italian lexemes were selected for the analysis in the online *Treccani* dictionary: *avvocato/avvocata* or *avvocatessa*, *architetto/architetta*, *medico/medica*, *autore/autrice*, *direttore/direttrice*, *cantante*, *professore/professoressa* or *professora*.²⁹

Only some of the female agentives analyzed are included in the dictionary as separate lemmas. It is unclear which criteria play a role here. For example, there is an entry for *medica*; furthermore, there is also an entry for *medichessa* with the indication that this term is used pejoratively or jokingly, while in other cases there is no separate entry for the feminine form at all. Due to the negative connotation that the suffix *-essa* had, especially in the past, the suffix *-a* is sometimes favored for female personal names (see Sabatini, 1987), even if the negative stigmatization no longer tends to be noticeable in usage. Lexems such as *professoressa* or *studentessa* can be used without hesitation (see Fusco, 2019, p. 37).

A look at the grammatical information on the lemmas reveals a rather inconsistent approach: the grammatical information for *avvocato*, for example, refers to the feminine form in brackets as follows: (f.: *-essa* o *-a*). However, this information is not available for all analyzed entries for male persons. Conversely, with some feminine lemmas there is a grammatical indication that it is a derivation of the masculine form, e.g., in *avvocata*: (femine form of *avvocato*). The feminine form *avvocatessa* is not included as a separate lemma in the dictionary. Still in the entry *avvocato* we find a reference to the negative connotation of the term, and the entry *avvocata* contains the information that *avvocatessa* is a synonym that is not in common use (without information on the negative connotation). In addition, *avvocata* does not refer to the professional title in the field of law but only to a meaning in the religious field (“*protettrice, interceditrice, attributo della Madonna o di sante*”³⁰).

In all entries included in the case study, the entry for the male person has the full scope, while the entry for the female person is very brief. The descriptions of the meanings of the male agentives often follow the person-neutral pattern “*chi predispone/professa/dirige/è ...*”³¹, while the definitions of the female agentives are usually introduced with “*donna che ...*”. Due to the brevity of the entries for female persons, they also contain very few examples or none at all.

It is worth mentioning the practice that towards the last part of the dictionary entries for male agentives there is often a separate reference to the female term, e.g., for *avvocato*: “*Per indicare una donna che esercita l’avvocatura nell’uso giuridico è stato usato il maschile avvocato, ma sono sempre più frequenti, nell’uso com., i femminili avvocatata e avvocatessa, quest’ultimo anche per indicare scherz. la moglie di un avvocato, o una donna che ha la parlantina sciolta, che si accalora nel discorrere e nel sostenere le ragioni proprie o altrui*”³² or *medico*: “*La forma femm. medichessa è dell’uso ant., o scherz. e spreg.; nell’uso attuale va diffondersi il femm. medica riferito*

²⁹ Eng. ‘lawyer’, ‘architect’, ‘physician’, ‘author’, ‘director’, ‘singer’, ‘professor’.

³⁰ “Patroness, intercessor, attribute of the Virgin Mary or saints” (translation by the authors).

³¹ Eng. ‘who prepares/provides/manages/is’.

³² See <https://www.treccani.it/vocabolario/avvocato/>.

a donna che eserciti la professione”.³³ The explanations seem to be formulated on an ad hoc basis with reference to the specific case. It remains unclear why the reference is missing in some cases, e.g., with *direttore*.

The extent to which automatically generated parts are included in the dictionary entries cannot be determined, so this aspect is excluded from the analysis.

4. State of the Art in Terminology Science³⁴

Before proceeding, some preliminary considerations are necessary. As stated in *TERMCAT* (2015, p. 4), lexicography organizes the study of language based on linguistic form: it identifies the canonical form and determines the different meanings that each form can take in real contexts. Therefore, it starts with the word to get to the meaning. Its approach, i.e., the semasiological approach, leads to treating masculine and feminine forms separately based on specific morphological criteria. Instead, terminology science deals with the study of concepts and their designations in specific domains (DIN 2342, 2011, p. 14 in Drewer & Schmitz, 2017, p. 61). The resulting activity, the terminology work, is based on an onomasiological approach. This approach identifies, firstly, the meaning and, secondly, the linguistic or non-linguistic designation(s)³⁵ assigned to the concept. Hence, in terminology science, it is not the linguistic form of the designation that is decisive but the concept, which, as such, is generally not related to gender (see *TERMCAT*, 2015, p. 4). In this regard, gender is considered a property, i.e., a qualitative attribute, and not a characteristic, i.e., an intrinsic attribute that forms and delimits concepts (Winter, 2021, p. 29; DIN 2342, 2022, p. 8; ISO 1087, 2019, 3.1.3).

Taking the aforementioned into consideration, how has the terminology world reacted to gender-sensitive and gender-neutral language? To answer this question, in recent years, Bengoechea (2017) analyzed *IATE*, the European Union’s inter-institutional terminology database. In particular, Bengoechea investigated how *IATE* dealt with feminine agentives. Her analyses resulted in the conclusion that “very few measures have been taken to implement non-sexist language in private or public terminology databases which include agentives” (Bengoechea, 2017, p. 215).

Drawing on this investigation, between 2020 and 2021, we also explored some terminology resources such as *IATE*, *TERMDAT*, *LinguaPC*, *TERMCAT*, *Euskalterm*, *FAOTERM*, *Termium Plus*®, and *Le Grand dictionnaire terminologique* (*GDT*), focusing on feminine agentives. The trend identified by Bengoechea seemed to be still current: in some databases, feminine agentives were not entered systematically if they were entered at all (Ralli & Evers, in review). For example, in *IATE* there

³³ See <https://www.treccani.it/vocabolario/medico3/>.

³⁴ In this section (4) and in the following sections on terminology databases (5.1 and 5.2), the agentives are translated into English exclusively when they are present in the respective databases. For the same reason, their definitions (5.1 and 5.2) are only given in the language provided in the respective databases.

³⁵ A concept can also be represented by definitions, formulas, symbols, graphics, etc.

was a matching entry³⁶ for the query *infirmiera*³⁷ with the corresponding form in German “Gesundheits- und Krankenpflegerin”, but there was no matching entry for *sindaca*³⁸. Other databases tried to treat systematically women and men on an equal footing by making “women linguistically visible besides men” (Motschenbacher, 2014 in Bengoechea, 2017, p. 206). That means feminine and masculine agentives were recorded in the same entry as independent items. This allows searching for female professionals, job titles, etc. via query mask.

This general analysis was the starting point for more in-depth research on our part. Different aspects needed to be closely analyzed:

- a. the presence of feminine agentives;
- b. how feminine agentives are recorded: whether gender is considered a property or a characteristic leads to different representations of the feminine agentives in terminology databases (Ralli & Evers, in review);
- c. to what extent feminine agentives are considered, i.e., only when they occur 1) as simple terms, e.g., *Richterin*³⁹ and 2) as complex terms, e.g., *nebenamtliche Richterin*⁴⁰, or also when they are part of complex terms, e.g., *auf Geheiss der Richterin*⁴¹;
- d. how the sexes are treated graphically (see Bengoechea, 2017, p. 217);
- e. how agentives are defined, i.e., whether a gender-sensitive and a gender-neutral language is used or not. Regarding this point, a clarification is necessary (Ralli & Evers, in review): in terminology work, concepts are generally defined by *genus proximum* and *differentiae specifica* (Magris, 1998, p. 41). Thus, the concept is described according to its intension, by stating its (closest) generic concept and specifying the delimiting characteristics (see Arntz et al., 2014, pp.63 ff.; ISO 704, 2022, 6.2). Our analysis focused on how and to what extent terminology databases can be gender-sensitive and gender-neutral, particularly when defining hyponyms.

To achieve these goals, we analyzed three well-known multilingual databases from Switzerland and Catalonia: *LinguaPC*, *TERMDAT*, and *TERMCAT*. We chose these databases because they proved to be good practice in our first investigation.

³⁶ <https://iate.europa.eu/entry/slideshow/1631775442783/1430984/it-de-en>, consulted on 15th September 2021.

³⁷ Feminine form of *nurse*.

³⁸ Feminine form of *mayor*.

³⁹ Feminine form of *judge*.

⁴⁰ Feminine form of *substitute judge*.

⁴¹ “By order of the judge”, where *Richterin* is the feminine form of *judge*.

5. Terminology Database Today

5.1 LinguaPC and TERMDAT

LinguaPC and *TERMDAT* are Swiss public terminology databases: the first is from Canton Bern and contains legal and administrative terms in German and French; the second is from the Federal Chancellery and includes the terminology of Swiss law, public administration, and the public sector in German, French, Italian, Romansh⁴², and, sometimes, in English⁴³. However, for space reasons, we focus exclusively on the German part of the terminological entries.

In both databases, German feminine and masculine agentives are treated as synonyms. For example, in *LinguaPC* the entry *Anwalt*⁴⁴ contains the corresponding feminine form *Anwältin*. Its synonyms *Rechtsanwalt* and *Fürsprecher* have the corresponding feminine forms *Rechtsanwältin* and *Fürsprecherin* too. The same applies in *TERMDAT*, where under *Anwalt*, *Anwältin*, *Fürsprecher*, and *Fürsprecherin* other synonyms are recorded: *Advokat* and *Advokatin*⁴⁵. Synonymity implies that all these agentives are lemmatized as separate items in the same terminological entry they belong to. That means in the query mask, you can start your search by entering the masculine or feminine form of an agentive, getting the same search results. Taking the above-mentioned example, you can search for both *Anwalt* and *Anwältin* or *Fürsprecher* and *Fürsprecherin*.

Masculine and feminine forms are also present in complex terms such as *eingetragener Anwalt* and *eingetragene Anwältin*. Also, in this case, the same rules apply as for the simple terms. The two investigated databases follow different strategies in complex terms, where an agentive is an object. *LinguaPC* tends to record both masculine and feminine agentives, such as *Patent eines bernischen Kindergärtners*⁴⁶ (m.) and *Patent einer bernischen Kindergärtnerin* (f.). However, this database does not contain enough complex terms of this type to confirm this assumption. On the contrary, *TERMDAT* tends to use the masculine agentive. For example, this is the case of *Ausschluss eines Miteigentümers* or *Abberufung eines Verwalters* where no feminine form (*Ausschluss einer Miteigentümerin*, *Abberufung einer Verwalterin*) is given in the terminological entry. The decision to omit the feminine form could have a methodological reason: in general, this database adopts a descriptive approach and documents designations as they are used in contexts. The presence of only masculine agentives may indicate a lack of authoritative and reliable sources attesting to the existence and use of the corresponding feminine agentives.

⁴² Romansh is a Rhaeto-Romance language spoken in the in the Swiss canton of the Grisons (Graubünden).

⁴³ <https://www.bk.admin.ch/bk/en/home/dokumentation/languages/termdat.html> (retrieved June, 12, 2024).

⁴⁴ Masculine form of *lawyer*.

⁴⁵ All the examples in this section are from *LinguaPC* (<https://www.linguaapc.apps.be.ch>) and/or *TERMDAT* (<https://www.termat.bk.admin.ch>), consulted until June 12, 2024. Wherever possible, the English translations of the terms used as examples for *LinguaPC* are sourced from *TERMDAT*.

⁴⁶ In this section, boldface has been added by the authors.

With regards to definitions, gender-neutral expressions (i.e., impersonal expressions, collective nouns) are generally used when a superordinate concept is defined. For example, *LinguaPC* defines *Anwalt* and *Anwältin* starting with the gender-neutral term *person*: “**Person**, die vom Obergericht das Anwaltspatent erhalten hat”. On the contrary, the definition of the hyponyms *amtlich bestellter Anwalt* or *eingetragener Anwalt* begins with a masculine designation, which represents their immediate superordinate concept, i.e., *Verteidiger*⁴⁷ and *Anwalt* respectively. The same pattern can be observed in other entries. For example, the definition of *Richter* and *Richterin*⁴⁸ is formulated as follows: “Vom Grossen Rat für eine Amtsdauer von sechs Jahren gewähltes **Mitglied** einer kantonalen Gerichtsbehörde”, where “Mitglied” corresponds to the gender-neutral term *member*. In the case of hyponyms like *ordentlicher Richter*⁴⁹ or *Einzelrichter*⁵⁰, *LinguaPC* defines them beginning with the masculine agentive *Richter* as its closest superordinate concept. *TERMDAT* applies the same approach: many agentives are defined using the masculine form or gender-neutral words.⁵¹ In addition, besides impersonal expressions or nouns, combined forms are also used here. For instance, *TERMDAT* defines *Bundesrichter* and *Bundesrichterin*⁵² as “**Richter/Richterin** am Schweizerischen Bundesgericht”. Considering this analysis, however not exhaustive, we can claim that much is being done in both databases to apply the principle of gender equality in terminology definitions, wherever possible.

5.2 TERMCAT

TERMCAT is a terminology consortium in charge of normalizing terminology in Catalan. Its mission includes, inter alia, the maintenance, and update of a Catalan terminology database with equivalent terms in other languages.⁵³

In its terminology database, masculine and feminine agentives in Catalan and Spanish are recorded together within the same field and separated by a glyph (“|”) as shown in the following example in Catalan⁵⁴:

1. advocat | advocada, advocadessa n m, f

This type of lemmatization violates two terminological principles (Ralli & Evers, in review): the first concerns term autonomy, according to which all terms in a terminological entry “are considered independent sub-units and can be described using

⁴⁷ Masculine form of *defence lawyer*.

⁴⁸ Masculine and feminine form of *judge*.

⁴⁹ Masculine form of *ordinary judge*.

⁵⁰ Masculine form of *single judge*.

⁵¹ See also the agentives *Richter* and *Richterin*, *unfähiger Richter* and *unfähige Richterin*, *Anwalt* and *Anwältin*, *Tierschutzanwalt* and *Tierschutzanwältin*, *Bundesanwalt* and *Bundeanwältin*, just to mention a few (retrieved June, 12, 2024).

⁵² Masculine and feminine form of *federal judge*.

⁵³ https://llengua.gencat.cat/en/direccio_general_politica_linguistica/02_organismes_vinculats/01_centre_de_terminologia_termcat/ (retrieved June, 13, 2024)

⁵⁴ All the examples of this section are from *TERMCAT*'s terminology database CERCATERM (https://www.termcat.cat/ca/cercaterm/advocata?type=basic&thematic_area=&language=) consulted before 13th June 2024.

the same set of data categories” (ISO 26162-1, 2019, 3.2.14); the second relates to data elementarity, according to which “a data field contains only one data element” (ISO 26162-1, 2019, 3.2.16). Nevertheless, it puts the masculine and feminine agentives at the same level, thus ensuring graphically equal treatment. Furthermore, this does not prevent searching for both forms via query mask, obtaining the same search results.

The same strategy applies both for simple terms and complex terms such as defense lawyer: *advocat defensor* (m.) and *advocata defensora* (f.) in Catalan or *abogado defensor* (m.) and *abogada defensora* (f.) in Spanish in procedural law. Regarding complex terms, no feminine form is provided where an agentive is an object.⁵⁵ This is the case of *petició d’advocat d’ofici*⁵⁶ in Catalan and *petición de abogado de oficio* in Spanish. The reason behind this choice may be the same as that described in Section 5.1. Finally, in the case of epicene designations, i.e., designations “that may refer to a male or a female”⁵⁷, like the Catalan word *pilot*⁵⁸, there is a double attribution of grammar, and the designation is entered once to avoid redundancy (TERMCAT, 2015, p. 3):

2. **pilot**, n m f⁵⁹.

Concerning definitions, these are only in Catalan as this is the pivot language of the database. As in the case of *LinguaPC* and *TERMDAT*, also *TERMCAT* tends to use gender-neutral terms for defining superordinate concepts. For example, the definition of *advocat*, *advocada*, and *advocadessa* begins with different gender-neutral terms according to the subject field the agentives belong to:

3. “**Persona** que assessora sobre qüestions de dret i defensa les causes civils, criminals o laborals d’altres.” (procedural law);
4. “**Professional** que es dedica a intervenir en causes o processos judicials i administratius, i a donar orientació i assessorament en matèria legal, tant a persones particulars com a professionals de l’àmbit de l’atenció a les persones” (welfare).

As is evident from these examples, also the delimiting characteristics are described with a gender-neutral language, too.

Also, in the case of subordinate concepts, we can observe similarities with the two Swiss databases. Their definitions usually begin with the designation of the respective superordinate concept, as in the entry *advocat matrimonialista* and *advocada matrimonialista*: “**Advocat especialista** en processos de separació i divorci”, where *advocat especialista* is the hyperonym of *advocat matrimonialista* in the masculine

⁵⁵ See as examples also *sol·licitud de designació d’advocat* (ca), *solicitud de designación de abogado* (es), “col·legi d’advocats” (ca), *colegio de abogados* (es) (retrieved June, 12, 2024).

⁵⁶ In this Section, boldface has been added by the authors.

⁵⁷ <https://www.collinsdictionary.com/de/worterbuch/englisch/epicene> (retrieved June, 12, 2024)

⁵⁸ Eng. ‘pilot’.

⁵⁹ https://www.termcat.cat/ca/cercaterm/pilot?type=advanced&thematic_area=&language=&condition=match&fields=denominacio&category=&hierarchy=&page=1 (retrieved June, 12, 2024).

form. Sometimes subordinate concepts are defined using gender-neutral designations or both masculine and feminine forms. This is the case of *advocat defensor* and *advocada defensora*⁶⁰:

5. “**Advocat o advocada** que defensa una persona que litiga, especialment en un procés penal” (procedural law).

At the time of Bengochea’s investigations (2017, p. 215) *TERMCAT* incorporated female agentives only for Catalan. Our analysis shows that now the Catalan database incorporates such agentives for Spanish, too. Like in *LinguaPC* and *TERMDAT*, respect for gender equality goes beyond the lemmatization of feminine agentives to the drafting of definitions in gender-neutral and gender-sensitive language. A basis for ensuring linguistic visibility and equality has been laid down.

6. Conclusion and Outlook

In German- and Italian-speaking countries respectively, *Duden* and *Treccani* are regarded as examples of good practice for publishers that have adopted new approaches to more gender-sensitive language use, and they did so in roughly the same period, i.e., 2021/2022. Therefore, the online versions of their dictionaries formed the basis for our case studies. Both pursue strategies to make female agentives more visible in their products. Both dictionaries have their own explicit information on the changing use of male and female agentives. In *Duden online*, this information, formulated according to a standardized pattern, can be found in separate info boxes on the website of each dictionary entry on male agentives. In *Treccani online*, the information is part of the dictionary entry (which closely follows the format familiar from print dictionaries) on masculine agentives.

In *Duden online*, female agentives are consistently lemmatized, followed by a complete dictionary entry. *Treccani online* is less consistent in this respect, so there is no separate entry for all feminine forms. It is noticeable in both dictionaries that the entries for female agentives are less extensive than those for male personal names, although the difference in *Treccani online* is more significant. There, women are relatively marginally visible, at least in the context of the case study.

Duden online uses a standardized definition practice, which usually refers to either a “männliche Person, die ...” or a “weibliche Person, die ...”. In *Treccani online*, this strategy is applied, albeit not consistently, to female agentives.

Some tendencies to perpetuate stereotypical role models are still recognizable. In *Duden online*, this applies in particular to the automatically generated typical word combinations. What is striking in *Treccani online* is the fact that a pejorative register marker often accompanies female personal names. This may be due to a peculiarity of the Italian language, but the dictionary certainly reflects this linguistic reality.

⁶⁰ See also *professor de formació viària* and *professora de formació viària* in the frame of road safety or *ajudant del jutge àrbitre* and *ajudanta del jutge àrbitre* (retrieved July, 31, 2024).

In the terminology field, *LinguaPC*, *TERMDAT*, and *TERMCAT* are among the databases that better apply gender equality in their data. From their representation, we could observe different underlying methodological choices. Considering our investigation, a common methodological approach is lacking, which would be relevant to facilitating data interchange.

We believe that many aspects need careful consideration, among them (Chiocchetti & Ralli, 2024; Ralli & Evers, in review):

- a. lemmatization of feminine agentives: we can have different types of representations depending on whether the feminine agentive is considered a property or a characteristic. If gender is considered a property, the feminine agentive is represented with the masculine agentive in the same entry. In this case, the following two scenarios are envisaged: the first, the feminine form is recorded as a synonym of the masculine agentive (see Section 5.1); the second, the feminine form is recorded together with the masculine agentive(s) within the same field (see Section 5.2). Instead, if gender is considered a characteristic, the feminine agentive is treated as a concept, thus in a dedicated entry and separated from the masculine agentive. In this case, a semasiological approach would be followed;
- b. term type and degree of syntagmatic complexity: feminine and masculine agentives can occur as simple terms (e.g., *Anwalt/Anwältin*), as complex terms (e.g., *eingetragener Anwalt/eingetragene Anwältin*), or as an object of a complex term with agentives as objects (e.g., *Patent eines bernischen Kindergärtners/Patent einer bernischen Kindergärtnerin*);
- c. drafting of definitions: the generic use of masculine agentives to refer to the closest generic concept (see Sections 5.1 and 5.2) can be avoided by combining several strategies,⁶¹ wherever possible. Some of them are also applied by *LinguaPC*, *TERMDAT*, and *TERMCAT*. However, as Ralli and Evers (in review) state: “the relationship between the hypernym and hyponym is semantic, i.e., linguistic, rather than conceptual because in this case the gender is treated as a property and not as a characteristic.”

The Council for German-Language Terminology, RaDT⁶², is working on all these aspects, among other things. A publication in this regard is planned soon.

Of course, we cannot consider our investigations exhaustive: more in-depth research is needed, also with regard to other languages and the impact of such representation on society. However, our analysis shows that it is high time for lexicographers and terminologists to deal with this topic systematically, studying shared standards for representing men and women more fairly and equally in lexical resources.

⁶¹ Such strategies should be applied according to each language’s grammatical typology (European Parliament 2018, p. 5).

⁶² Rat für deutschsprachige Terminologie, <http://radt.org/>

References

- Arntz, R., Heribert, P., & Schmitz, K-D. (2014). *Einführung in die Terminologearbeit*. 7. OLMS.
- Bazzanella, C. (2009). Stereotipi e categorizzazioni del femminile/maschile. In S. Regazzoni (Ed.), *Mi fai male... Atti del Convegno Venezia, Auditorium Santa Margherita 18-19-20 novembre 2008*. Cafoscarina.
- Bengoechea, M. (2017). Agentives for women, a gap still to fill: dismissing non-sexist language policies in terminological resources. In P. Faini (Ed.), *Terminological approaches in the European context* (pp. 200–221). Cambridge Scholars Publishing.
- Calvo Ramos, L. (1998). Algunos rasgos sexistas en el Diccionario de la Real Academia Española de la lengua. *Revista Llingua i Dret*, 30, 47–62.
- Cavagnoli, S., & Dragotto, F. (2021). *Sessismo*. Mondadori Università.
- Chiocchetti E, & Ralli, N. (in review). Comunicazione istituzionale inclusiva in Alto Adige: esperienze e questioni aperte. *Convegno Internazionale di Linguistica e Glottodidattica Italiana CILGI5 2023*. March 23-25, 2023, Salerno.
- De Bernardis, I. (2007). Reviewed Work(s): Piccolo dizionario dell'inuguaglianza femminile by Alice Ceresa and Jacqueline Risset. *Belfagor*, 62(5), 616–619.
- DIN 2342 (2022). *Terminologiewissenschaft und Terminologearbeit—Begriffe*. Beuth.
- Drewer, P., & Schmitz, K.-D. (2017). *Terminologiemanagement. Grundlagen – Methoden – Werkzeuge (Kommunikation und Medienmanagement)*. Springer Vieweg.
- European Parliament (2018). *Gender-neutral language in the European Parliament*. Retrieved June 15, 2024, from https://www.europarl.europa.eu/cmsdata/151780/GNL_Guidelines_EN.pdf
- Euskalterm: *Banco Terminológico Público Vasco*. Retrieved September 18, 2020, from https://www.euskadi.eus/gobierno-vasco/contenidos/informacion/euskalterm/es_7553/euskalterm.html
- Evers, E. (2022). Geschlechtersensible Sprache — wie bringe Ich es meiner Terminologiedatenbank bei? *edition 2/22*, 13–20.
- FAOterm: *FAO Terminology Portal*. Retrieved September 18, 2020, from <https://www.fao.org/faoterm/en/>
- Fuertes-Olivera, P. A., & Tarp, S. (2022). Critical Lexicography at Work: Reflections and Proposals for Eliminating Gender Bias in General Dictionaries of Spanish. *Lexikos*, 32(2), 105–132.
- Fusco, F. (2009). Stereotipo e genere: il punto di vista della lessicografia. *Linguistica*, 49(1), 205–225. doi:10.4312/linguistica.49.1.205-225
- Fusco, F. (2019). Il genere femminile tra norma e uso nella lingua italiana: qualche riflessione. In S. Adamo, G. Zanfabro, & E. Tigani Sava (Eds.), *Non esiste solo il maschile. Teorie e pratiche per un linguaggio non discriminatorio da un punto di vista di genere* (pp. 27–49). EUT. Edizioni Università di Trieste.

- Graham, A. (1975). The making of a non-sexist dictionary. In B. Thorne, & N. Henley (Eds.), *Language and Sex: Difference and Dominance* (pp. 57–63). Newbury House Publishers.
- Hampares, K. J. (1976). Sexism in Spanish Lexicography? *Hispania*, 59(1), 100–109. doi:10.2307/339381
- Hanks, P. (2012). The Corpus Revolution in Lexicography. *International Journal of Lexicography*, 25(4), 398–436. <https://doi.org/10.1093/ijl/ecs026>
- IATE: *Interactive Terminology for Europe*. Retrieved June 12, 2024, from <https://iate.europa.eu/>
- ISO 1087 (2019). *Terminology work and terminology science—Vocabulary*. ISO.
- ISO 704 (2022). *Terminology work—Principles and methods*. ISO.
- Kotthoff, H., Nübling, D., & unter Mitarbeit von Schmidt, C. (2018). *Genderlinguistik: eine Einführung in Sprache, Gespräch und Geschlecht*. Narr Francke Attempto.
- Le Grand dictionnaire terminologique (GDT). Retrieved September 18, 2020, from <https://vitrinelinguistique.oqlf.gouv.qc.ca/> (last consulted on 18th September 2020)
- Lepschy, A. L., Lepschy, G., & Sanson, H. (2001). Lingua italiana e femminile. *Quaderns d'Italia* 6, 9. doi:10.5565/rev/qdi.51
- Lepschy, G. (1987). Sexism and the Italian language. *The Italianist*, 7(1), 158–169. doi:10.1179/ita.1987.7.1.158
- Lingua-PC: *Terminologiedatenbank des Kantons Bern*. Retrieved June 12, 2024, from <https://www.linguapc.apps.be.ch/>
- Magris, M. (1998). *La definizione in terminologia e nella traduzione specialistica*. Rivista internazionale di tecnica della traduzione. *International Journal of Translation*, 3, 37–63.
- Manera, M., & Bazzanella, C. (2006). Gender on-line in the Italian Word Thesaurus. In M. Hellinger, & H. Bußmann (Eds.), *Gender, language and new literacy* (pp. 107–122). Continuum.
- Müller-Spitzer, C., & Lobin, H. (2022). Leben, lieben, leiden: Geschlechterstereotype in Wörterbüchern, Einfluss der Korpusgrundlage und Abbild der sprachlichen, Wirklichkeit. In G. Diewald, & D. Nübling (Eds.), *Genus – Sexus – Gender* (pp. 33–649).
- Müller-Spitzer, C. (2022). Zumutung, Herausforderung, Notwendigkeit? Zum Stand der Forschung zu geschlechtergerechter Sprache. *Geschlechtergerechte Sprache*, 23–29.
- Nübling, D. (2009). *Zur lexikografischen Inszenierung von Geschlecht. Ein Streifzug durch die Einträge von Frau und Mann in neueren Wörterbüchern* 37(3), 593–633. <https://doi.org/doi:10.1515/ZGL.2009.037>
- Porsch, P. (2004). Frau im Wörterbuch – Das DUDEN-Universalwörterbuch 2003 als Fortsetzung eines Trivialromans. In U. Fix, M. Schröder, & H. Wellmann (Eds.), *Zwischen Lexikon und Text – lexikalische, stilistische und textlinguistische Aspekte* (pp. 358–365). Verlag der Sächsischen Akademie der Wissenschaften zu Leipzig.

Ralli N., & Evers, E. (in review). To gender or not to gender, that is the question: gender-inclusive language in the legal context. EAFT Summit 2023.

Sabatini, A. (1987): *Il sessismo nella lingua italiana*. Commissione Nazionale per la parità e le pari opportunità tra uomo e donna.

TERMCAT (2015): *Denominació: Representació del femení (particularitats)*. TERMCAT.

TERMCAT: *Cercaterm*. Retrieved June 12, 2024, from <https://www.termcat.cat/ca>

TERMDAT: *The Federal Administration's terminology database*. Retrieved June 12, 2024, from <https://www.bk.admin.ch/bk/en/home/dokumentation/languages/termdat.html>

Termium Plus®: *the Government of Canada's terminology and linguistic data bank*. Retrieved June 12, 2024, from <https://www.btb.termiumplus.gc.ca/tpv2alpha/alpha-eng.html?lang=eng>

Thüne, E.-M., & Leonardi, S. (2006). The German Word Thesaurus and socio-cultural models. In E.-M. Thüne, S. Leonardi, & C. Bazzanella (Eds.), *Gender, Language and New Literacy: a Multilingual Analysis*. (pp. 79–94). Continuum.

Westveer, T., Sleeman, P., & Aboh, E. O. (2018). Discriminating dictionaries? feminine forms of profession nouns in dictionaries of French and German. *International Journal of Lexicography* 31(4), 371–383.

Whitcut, J. (1984). Sexism in dictionaries. In R. R. K. Hartmann (Ed.), *LEXeter '83: proceedings* (pp. 141–144). De Gruyter. doi:10.1515/9783111593166-020

Winter T. (2021). Das Gendern. Terminologisch betrachtet ein Missverständnis mit Skandalpotential. *edition 2/21*, 29-30.

Acknowledgements

We kindly thank experts from the Central Language Services, Terminology Section of the Swiss Federal Chancellery, and the Central Service of Terminology of the State Chancellery of the Canton of Bern for their valuable input. We would also like to thank our colleague Flavia De Camillis for working with us on the study.

Contact information

Andrea Abel

Free University of Bozen-Bolzano – Faculty of Education /
Eurac Research – Institute for Applied Linguistics
andrea.abel@unibz.it
andrea.abel@eurac.edu

Natascia Ralli

Eurac Research – Institute for Applied Linguistics
natascia.ralli@eurac.edu

Ieda Maria Alves, Beatriz Curti-Contessoto, and
Ana Maria Ribeiro de Jesus

CHALLENGES OF CREATING A MEDICAL DICTIONARY FOR A LOW LITERACY AUDIENCE IN BRAZIL

Focusing on Politically Marked Terms
Related to the COVID-19 Pandemic

Abstract The COVID-19 pandemic has impacted numerous sectors at different levels and has imposed a radical change in the pace of life in societies across the globe. A partially technical vocabulary related to COVID-19 quickly became part of everyday life, introduced mainly by news and official bodies. To describe the characteristics of the terminology being disseminated in Brazil, the project *Study and dissemination of COVID-19 terminology (2020–2024)* was proposed. More specifically, it has been possible to observe the recurrence of lexical units closely related to political issues – which is the focus of this paper. Entries like *gripezinha* (small flu), *tratamento precoce* (early treatment), *kit covid* (COVID kit) and *ministério paralelo* or *gabinete paralelo* (parallel ministry or parallel office) will be explored in this paper. These examples show cases of politically marked terms related to the COVID-19 Brazilian pandemic context. In Brazil, the pandemic has been marked by science denialism, especially on the part of the federal government. Due to this political influence, particularly concerning interference in the Brazilian public health system, these terms, which are not officially recognized by the Brazilian and international medical community, were included in the proposed dictionary, supplemented with contextual information elucidating the political dimensions entailed.

Keywords COVID-19; science denialism; political interference

1. Introduction

The COVID-19 pandemic has had a profound impact on numerous sectors at various levels, fundamentally altering the pace of life in societies across the globe. This unprecedented health crisis not only challenged healthcare systems but also brought about significant changes in economic activities, education, and social interactions. As countries grappled with lockdowns, social distancing measures, and travel restrictions, the way people worked, learned, and connected with others underwent a dramatic transformation.

A notable aspect of this period has been the rapid integration of a partially technical vocabulary related to COVID-19 into everyday life. This vocabulary was primarily introduced through news and official communications from health authorities and government institutions, aiming to inform and guide the public in understanding and combating the virus. Thus, specialized medical terms, such as *comorbidities* and *strain*, became widely known and used by the general public.

To describe the characteristics of the terminology being disseminated in Brazil, based on the principles of communicative theory of terminology (Cabr , 1999), the project *Study and dissemination of COVID-19 terminology* (2020–2024) was proposed. In the scope of this project, a terminological dictionary aimed at non-specialised speakers in the medical field and with little formal education is being elaborated. To achieve this, we collected texts from Brazilian institutions (Butant  Institute, Fiocruz, Fapesp) and international sources (World Health Organization (WHO), Pan American Health Organization (PAHO)) – official corpus (OC) – along with texts extracted from widely circulated Brazilian newspapers (*Folha de S. Paulo*, *O Globo*, *O Estado de S o Paulo*) – journalistic corpus (JC) – from the year 2020 onward. The motivation for the creation of this dictionary arises from the fact that 29% of the Brazilian population has difficulty in interpreting texts and performing simple mathematical operations in their daily activities (INAF, 2018). Because of that, all entries were prepared by following the principles of an international tendency towards using plain language (*linguagem simples*), more easily understood by users who are not specialised in the area in question. Following these principles, the definitions and explanatory or complementary notes, in relation to the term, are being written in *plain language*, which designates texts understandable by different types of speakers (Gasperin, Maziero, & Alu sio, 2010; Cortina, Delgado, & Finatto, 2021; ISO 24495-1, 2023).

The corpus collected for this project enabled the observation of terms that extend beyond specialized medical knowledge and reveal a political bias. The study presented in this paper details the results of a descriptive analysis of these specific terms, linking them to the political landscape of Brazil during COVID-19 pandemic. It also outlines how these terms were documented in the proposed dictionary.

Considering these objectives, this paper is organized as follows: after this introduction, section 2 presents the Brazilian political context during the COVID-19 pandemic, listing relevant facts and linguistic expressions associated to them. Section 3 discusses the decisions related to the inclusion of politically marked terms in the conceptual system created for the project and provides examples of the entries in the proposed dictionary. Some final considerations are made in the last section.

2. Brazilian Political Context

The health crisis triggered by the spread of the virus causing COVID-19 has required efforts from various sectors of societies worldwide to combat its spread and minimize its effects. Preventive measures such as social distancing, frequent hand hygiene with 70% alcohol gel or soap and water, and the use of masks have been recommended by international organizations like WHO and PAHO.

Brazil’s political response to these recommendations, however, was marked by scientific denialism on the part of the federal government. It offered insufficient transparency in reporting statistics and provided conflicting information on COVID-19. It also neglected scientific information, acting against social distancing and health measures

determined by states and municipalities, and delaying vaccination and financial support distribution (cf. De Almeida et al, 2022).

For instance, on March 24th, 2020, the President of the Republic, Jair Bolsonaro, transmitted an official statement in which he attacked the measures such as social distance and lockdown. On that occasion, as many times, he questioned the efficacy of the vaccines and declared that the citizens should return to their normal activities and the students should return to school. During the former president's speech, there were several *panelaços* (pot protests) in different Brazilian cities, demonstrating that a portion of the population disagreed with the federal government's position (cf. UOL, 2020).

In his speeches, former President Jair Bolsonaro often downplayed the importance of the pandemic. He would sometimes refer to COVID-19 as *gripezinha* (small flu), clearly intending to emphasize that the symptoms were not severe (cf. O Globo, 2020).

This scientific denialism also influenced guidelines from the Brazilian Ministry of Health at the time, which lacked scientific basis, such as the promotion of *tratamento precoce* (early treatment) (cf. Brazilian Ministry of Health, 2020). To initiate early treatment – term employed to designate treatments administered to individuals who are not yet infected with the new coronavirus – Brazilian citizens were encouraged to use the *kit COVID* (COVID kit), which included “drugs such as hydroxychloroquine, ivermectin, nitazoxanide, azithromycin, and systemic corticosteroids” (JUSP, 2021) and was aimed at preventing contagion by the new coronavirus or alleviating its symptoms.

The use of the COVID kit was encouraged by the Federal Council of Medicine and the Ministry of Health, with explicit support from the federal government. However, this encouragement went against the recommendations of national medical and scientific organizations, such as the Brazilian Medical Association (AMB) and the Brazilian Societies of Infectious Diseases and Pulmonology and Phthisiology, as well as international organizations like the National Institutes of Health (NIH) and the Infectious Diseases Society of America (IDSA). These organizations advised against the use of the COVID kit, citing the lack of proven clinical benefits and the potential harm these drugs could cause to patients. Despite this, federal authorities promoted the use of the COVID kit, leading to its distribution by health insurance companies and municipalities to their doctors for prescription to patients (JUSP, 2021).

Given the repercussions of this scientific denialism, in 2021, the then Minister of the Federal Supreme Court established a Parliamentary Commission of Inquiry (CPI), known as the *CPI da COVID-19* (COVID-19 CPI), to investigate the possible failures of the Federal Government in dealing with the pandemic, including identifying those potentially responsible. This CPI revealed the existence of a *ministério paralelo* (parallel ministry), which was a group that acted as advisors to the former president Jair Bolsonaro outside the Ministry of Health. Political influencers promoted the so-called early treatment in collaboration with Jair Bolsonaro's government, while a company – Prevent Senior – was responsible for validating the effectiveness of this treatment through purported studies (Folha de S. Paulo, 2021).

The report of the CPI concluded that Bolsonaro, “supported by the parallel ministry defended the herd immunity through collective contamination by the virus”. As the federal strategy was stimulating the spread of the coronavirus, Bolsonaro’s speeches were against the adoption of preventive measures – especially social distance and the use of masks – to avoid contamination. This report states that:

The federal government, repeatedly, encouraged the Brazilian population to carry on with their lives normally, without taking any possible precautions. To defend this point of view, it used to invoke the protection and preservation of the economy and to encourage the maintenance of any and all economic activities, as well as face-to-face classes in public and private education. Advertising campaigns focusing on the economy were carried out despite health risks (Comissão Parlamentar de Inquérito do Senado, 2021).

From a medical perspective, the concept of herd immunity would be important once a vaccine is available. Therefore, if at the time when the federal government was advocating this strategy, an effective vaccine had existed and a large portion of the Brazilian population had been vaccinated, herd immunity could have been viable. However, the CPI report shows that this was not the case because there were no vaccines available at that time. Exposing the population to the virus, as advocated by the then federal government in Brazil, was a strategy that could put people’s lives at risk and overload the health system. For these reasons, the term *imunidade de rebanho* (herd immunity), previously unfamiliar to the general Brazilian public, began to acquire a negative connotation in this political context. Implementing this strategy without scientific backing and without vaccines effectively suggested that lives were expendable in the name of protecting the economy (UFSM, 2020; Butantan, 2020).

3. Lexicographic Decisions Related to Politically Marked Terms

The terms highlighted in the previous section are politically marked in Brazil and are not strictly medical terms recognized by the Brazilian and international medical communities as related to the COVID-19 pandemic. However, their significant circulation in Brazil led to their inclusion in the nomenclature of the proposed dictionary. Thus, following Cabré’s (2008) precepts on the polyhedral nature of terms, which include linguistic, cognitive, and discursive dimensions, the expressions discussed in this paper can be considered terms within the political and social contexts in which they were disseminated.

In this dictionary, the terms and their respective entries are presented by following an onomasiological system. This structure was established according to the DeCS/MeSH (Health Sciences Descriptors, 2021), released by the PAHO. These descriptors consider categories such as causative agent, anatomy, prevention, diagnosis, disease, treatment, government, health policies, among others.

Specifically, regarding the terms covered in this paper, a challenge has arisen: since they are not strictly medical terms, into which categories should these terms be included within the scope of our study?

The term *gripezinha* was used by the former president to refer to the disease COVID-19. It would be logical to think that it should therefore be included in the disease category of the dictionary's conceptual system. However, this inclusion could lead the reader to make an erroneous association between the seriousness of the disease and the politically marked expression.

In turn, the term *tratamento precoce*, whose syntagmatic structure has *tratamento* as its noun nucleus, seems to indicate a specific conceptual relation with other terms in the generic category named treatment. However, this relation is not correct from a medical point of view, since early treatment, as we have seen, is not a type of treatment that has scientific backing.

In a similar vein is the term *kit COVID*. Used in the Brazilian political context to refer to a set of drugs that supposedly promote early treatment, this term could fall under the treatment category. Nevertheless, conceptually associating it with scientifically proven methods of treating COVID-19 would be a mistake.

The term *ministério paralelo* (and its variant *gabinete paralelo*) might suggest a direct conceptual relation with the government organizations combating COVID-19 in Brazil. However, their actions were entirely contrary: this "ministry" existed irregularly to support the scientific denialism propagated by the Brazilian federal government at the time.

Of all these examples, the term *imunidade de rebanho* (or *imunidade coletiva*) is the only one that existed before the COVID-19 pandemic. It is, therefore, a term officially related to the field of health and could, in principle, be included in the dissemination category. However, in Brazil, it was used by the former president to justify the federal government's strategy of resisting preventive measures. The use of this term in his political speeches gave it a negative connotation, associating it with a disregard for the seriousness of the disease. For this reason, it should not be officially linked to the COVID-19 pandemic, as argued by Brazilian medical experts at the time.

The intrinsic connection between these terms and Brazilian politics during the pandemic led us to include them in the government category. This category, provided by DeCS/MeSH, aligns with the criteria we adopted for organizing the onomasiological system in our work.

In this way, the political interference in the dissemination of information related to COVID-19 in Brazil is evident and is also present in the dictionary's microstructure, which is composed by: entry, grammatical information, definition, contextual information and, if necessary, a note. For illustration, we present the entries *tratamento precoce* and *kit COVID*, which are included under government category. The entry is presented in Portuguese, followed by its English translation.

tratamento precoce s.m.

En early treatment

Tratamento de prevenção com a utilização de medicamentos que os estudos científicos mostram não serem eficazes contra a Covid-19.

O Ministério da Saúde divulgou novas orientações para uso de medicamentos, mantendo o uso da cloroquina ou da hidroxicloroquina, no tratamento precoce de pacientes com Covid-19, no Sistema Único de Saúde (SUS). A partir de agora, foi ampliado o uso desses medicamentos para gestantes e crianças e adolescentes, que passaram a fazer parte dos grupos de risco. Esses medicamentos são indicados para casos leves, moderados e graves. (<OC_MinSaude_2020-06-18>)

Nota: Incentivado pelo governo federal, esse tratamento recomenda o uso dos medicamentos cloroquina, hidroxicloroquina, ivermectina, azitromicina, nitazoxanida, dentre outros, não recomendados por médicos contra a Covid-19.

early treatment n.

Preventive treatment with drugs that scientific studies have shown to be ineffective against COVID-19.

The Ministry of Health has released new guidelines for the use of medicines, continuing the use of chloroquine and hydroxychloroquine for the early treatment of COVID-19 patients in the Unified Health System (SUS). The use of these drugs has now been extended to pregnant women, children, and adolescents, who are considered part of the risk groups. These drugs are recommended for mild, moderate, and severe cases. (<OC_MinSaude_2020-06-18>)

Note: Encouraged by the federal government, this treatment recommends the use of drugs such as chloroquine, hydroxychloroquine, ivermectin, azithromycin, and nitazoxanide, among others, which are not recommended by doctors for treating COVID-19.

Kit COVID s.m.

En COVID kit

Conjunto de medicamentos recomendados pelo governo do ex-presidente Jair Bolsonaro para evitar o contágio pelo novo coronavírus ou amenizar os sintomas da COVID-19.

Pesquisadores da USP relatam em artigo na revista científica *The Lancet Regional Health – Americas* como ocorreu a disseminação do “tratamento precoce da covid-19” com o chamado “kit covid” no Brasil. Referendado –

sem comprovação científica – por autoridades públicas e médicos e bastante difundido nas redes sociais, o “kit covid” inclui drogas como hidroxicloroquina, ivermectina, nitazoxanida, azitromicina e corticosteroides sistêmicos, e tem sido amplamente utilizado no País porque, supostamente, poderia tratar a covid-19 de forma precoce e, assim, evitar hospitalizações e mortes. (<JC_JUSP_2021-10-14>)

COVID kit n.

Set of medications recommended by the government of former President Jair Bolsonaro to prevent the spread of the new coronavirus or to alleviate the symptoms of COVID-19.

Researchers from the University of São Paulo reported in an article in the scientific journal *The Lancet Regional Health – Americas* how the dissemination of the “early treatment of COVID-19” with the so-called “COVID kit” occurred in Brazil. As it was endorsed – without scientific proof – by public authorities and doctors and widely spread on social media, the “COVID kit” includes drugs such as hydroxychloroquine, ivermectin, nitazoxanide, azithromycin, and systemic corticosteroids, and it has been widely used in the country because it supposedly could provide an early treatment of COVID-19 and thus prevent hospitalizations and deaths. (<JC_JUSP_2021-10-14>)

As demonstrated by these examples, the definitions of *tratamento precoce* and *kit COVID* not only clarify the meaning of these terms in the specific context of COVID-19 but also highlight particular political aspects in Brazil. The contextual information presented support this connection between COVID-19 and Brazilian politics, and the notes, where provided, elucidate other pertinent elements.

4. Final Remarks

This paper presents a cross-section of the results achieved by the Project *Study and dissemination of COVID-19 terminology*, focusing on politically marked terms in Brazilian Portuguese identified in the studied corpora. To elucidate the link between the dissemination of these terms and politics in Brazil, this paper provides a brief contextualization of the federal government’s strategies during the COVID-19 pandemic.

The facts elucidate that the approach adopted by former President Jair Bolsonaro and his political supporters, backed by the guidelines from the Ministry of Health, contradicted the recommendations put forth by scientists and health experts in Brazil and the international community. This concerted effort in favor of interests, particularly economic interests, by the federal government, propagated scientific denialism throughout the country.

Indeed, a portion of the population did adhere to the federal recommendations,

including seeking early treatment. This situation in Brazil underscores the profound disconnection between science and a part of Brazilian society.

Political decisions influenced the handling of the COVID-19 pandemic, resulting in severe health repercussions and endangering lives, which left an imprint on the lexicon associated with the pandemic. Consequently, the proposed dictionary not only documents linguistic data regarding these terms, accentuating their inherent political dimensions, but also preserves this part of Brazil's history during this dark period for humanity.

References

- Brazilian Ministry of Health. (2020). *Ministério da Saúde reforça ação pelo enfrentamento à Covid-19 na região Sul*. Retrieved February 7, 2024, from <https://www.gov.br/saude/pt-br/assuntos/noticias/2020/julho/ministerio-da-saude-reforca-acao-pelo-enfrentamento-a-covid-19-na-regiao-sul>
- Butantan. (2020). *O que é imunidade de rebanho?* Retrieved March 7, 2024, from <https://coronavirus.butantan.gov.br/ultimas-noticias/o-que-e-imunidade-de-rebanho>
- Cabré, M. T. (1999). *La terminología: representación y comunicación. Elementos para una teoría de base comunicativa y otros artículos*. Institut universitari de lingüística aplicada (IULA).
- Cabré, M. T. (2008). El principio de poliedricidad: la articulación de lo discursivo, lo cognitivo y lo lingüístico en Terminología. *Ibérica*, 16, 9–36.
- Comissão Parlamentar de Inquérito do Senado. (2021). *CPI da Pandemia*. Retrieved May 30, 2024, from <https://legis.senado.leg.br/comissoes/comissao?codcol=2441>
- Cortina Silva, A. F., Delgado, H. O. K., & Finatto, M. J. (2021). Acessibilidade textual e terminológica para o português brasileiro: pesquisa, estratégias e orientações de [re]escrita. *Revista Moara*, 58, 322–343.
- De Almeida, L., Carelli, P. V., Cavalcanti, N. G., do Nascimento J.-D. Jr, & Felinto, D. (2022). Quantifying political influence on COVID-19 fatality in Brazil. *PLoS ONE* 17(7), e0264293. <https://doi.org/10.1371/journal.pone.0264293>
- Folha de S. Paulo. (2021). *CPI avalia que caso Prevent pode chegar ao governo e traz gabinete paralelo de volta ao foco*. Retrieved May 30, 2024, from <https://www1.folha.uol.com.br/poder/2021/09/cpi-avalia-que-caso-prevent-pode-chegar-ao-governo-e-traz-gabinete-paralelo-de-volta-ao-foco.shtml>
- Gasperin, C., Maziero, E., & Aluísio, S. M. (2010). Challenging choices for text simplification. In A. Branco, A. Klautau, R. Vieira, & Vera L. Strube de Lima (Eds.), *Proceedings of PROPOR 2010* (pp. 40–50), 9th International Conference (PROPOR), Springer.
- Health Sciences Descriptors. (2021). *DeCS – Descritores em Ciência da Saúde*. Retrieved May 30, 2024, from <https://decs.bvsalud.org/en/>

INAF. (2018). *Indicador de analfabetismo funcional: Em 20 anos, índice de analfabetos funcionais caiu de 40% para 30%*. Retrieved May 30, 2024, from <https://alfabetismofuncional.org.br/>

ISO 24495-1. *Plain language — Part 1: Governing principles and guidelines*. Geneva: ISO, 2023.

JUSP. (2021). “*Tratamento precoce*” e “*kit covid*”: a lamentável história do combate à pandemia no Brasil. Retrieved May 30, 2024, from <https://jornal.usp.br/ciencias/tratamento-precoce-e-kit-covid-a-lamentavel-historia-do-combate-a-pandemia-no-brasil/>

O Globo. (2020). *Bolsonaro volta a minimizar pandemia e chama Covid-19 de ‘gripezinha’*. Retrieved May 30, 2024, from <https://oglobo.globo.com/politica/bolsonaro-volta-minimizar-pandemia-chama-covid-19-de-gripezinha-1-24319177>

UFSM. (2020). *O que é “imunidade de rebanho” e por que não há consenso científico sobre sua eficácia?*. Retrieved May 30, 2024, from <https://www.ufsm.br/midias/experimental/agencia-da-hora/2020/07/24/o-que-e-imunidade-de-rebanho-e-porque-nao-ha-consenso-cientifico-sobre-sua-eficacia>

UOL. (2020). *‘Gripezinha’: leia a íntegra do pronunciamento de Bolsonaro sobre covid-19*. Retrieved May 30, 2024, from <https://noticias.uol.com.br/politica/ultimas-noticias/2020/03/24/leia-o-pronunciamento-do-presidente-jair-bolsonaro-na-integra.htm?cmpid=copiaecola>

Contact information

Ieda Maria Alves

Universidade de São Paulo
iemalves@usp.br

Beatriz Curti-Contessoto

Universidade de São Paulo
bfcurti@gmail.com

Ana Maria Ribeiro de Jesus

Universidade Federal do Espírito Santo
ana.m.jesus@ufes.br

Lynne Bowker

EPONYMS, EDI AND TERMINOLOGY PLANNING IN THE BIOLOGICAL SCIENCES

Abstract Traditional practices for naming species in the biological sciences often incorporate eponyms. However, the group of honourees is not very diverse, and many individuals have links to colonization. A grassroots movement is emerging within the biological sciences to give new scientific and/or common names to species that bear harmful eponyms. Approaches to renaming species include updating terminology planning processes, using more diverse and inclusive eponyms, re-instating pre-colonial names, and replacing eponyms with transparent terms. Many of these activities are in collaboration with Indigenous communities, as well as with other types of experts and the broader public. It is important for terminologists to be aware of these developments because they are well positioned to contribute to such discussions moving forward.

Keywords binomial nomenclature; common names; eponyms; retro-terminologization; scientific names; terminology planning; transparency

1. Introduction

Recent years have seen an increased awareness across multiple sectors of society about the particular damage caused by harmful naming practices related to settler colonialism, and renaming buildings, streets, awards, etc., can be viewed as part of a broader movement to increase social justice and to improve various dimensions of equity, diversity, and inclusion (EDI).

While names are distinct from terms (the former designate individual objects or people and the latter designate concepts or general notions), re-designation is relevant to Terminology. Terminology is the study of and the field concerned with the collection, description processing and presentation of terms (i.e., lexical items belonging to specialized areas of usage of one or more languages) (Sager, 1990). As noted by Sager (1990, p. 62), “most new terms are formed as and when new concepts are created in such circumstances as new discoveries, restructuring of existing knowledge, incidental observations or planned industrial developments,” meaning that terms are often chosen by scientists rather than by terminologists. As a result, “the linguistic sign for a concept can be quite arbitrarily chosen and often is” (Sager, 1990, p. 62).

In the biological sciences, scientists regularly use eponyms when naming species. In hindsight, it is clear that this practice has perpetuated harm, such as when the person honoured is associated with colonialism. There is now a grassroots movement to redress this situation, and this article will explore the re-designation practices that are emerging in the biological sciences to replace undesirable eponyms. The article first outlines how eponymous terms can enter specialized communication

and why this can be problematic. Next, the focus moves specifically to eponymous naming practices in the biological sciences and the resulting impact. Finally, several approaches for rectifying the use of eponyms are considered.

While much of the naming and re-designation of species in the biological sciences has been carried out by scientists rather than terminologists, terminologists need to be aware of these activities since many species have a common name in the local language in addition to a scientific name. Terminologists may be asked to support terminology planning for these common names, so they must understand the approaches that are being employed, and know whether or not these appear to be successful. Such insights can help to inform future EDI-oriented terminology work, whether in the biological sciences or in other domains.

2. Eponyms Used in Specialized Communication

Historically, a common method of term formation is the eponymic compound, which combines the name of a person or place with a substance, material, object, instrument, method, process, measure, etc. (Sager, 1990, p. 77). This practice has led to many inventions being named after their inventors (e.g., *Faraday cage*, *Ferris wheel*, *Phillips screwdriver*, *Rorschach test*). One drawback of eponymic compounds is that they are not transparent because they do not express the essential characteristics of the concept that they designate. This makes it hard for people to understand the nature of the concept simply by looking at the term. For instance, compare the limited information conveyed by the eponymic compound *Phillips screwdriver* compared to its non-eponymous alternative *cross-head screwdriver*.

Beyond lack of transparency, eponymic naming practices can also create EDI issues. For instance, as pointed out by Enserink (2013), “Historically, many infectious disease agents — or the diseases themselves — have been named after the place where they were first found. Increasingly, however, scientists and public health officials have shied away from that system to avoid stigmatizing a particular country or city.” Such stigmatization led to racism towards Asian people during the COVID-19 pandemic, when informal terms such as *China flu* or *Wu Flu* (derived from the name of the city of Wuhan) emerged to refer to COVID-19 (Lee & Johnstone, 2021). Selecting a non-stigmatizing name for the novel coronavirus was a priority for the International Committee on Taxonomy of Viruses (ICTV), who moved quickly to establish the terms *SARS-CoV-2* (to refer to the virus) and *COVID-19* (to refer to the disease caused by the virus) before the stigmatizing alternatives could take hold (Gorbalenya et al., 2020). This lesson was learned the hard way through previous instances where a stigmatizing term could not be dislodged, such as in the cases of Middle East Respiratory Syndrome (MERS) and the Spanish Flu (Lee & Johnstone, 2021). Although it can be challenging to dislodge established terms, it is not impossible. At present, a grassroots EDI-motivated effort is gaining traction within the biological science community to rename species whose eponymic references are considered harmful because they are linked to colonialism.

2.1 Traditional Approach to Naming Species

In the early days of naming plants and animals, the focus was on designating local flora and fauna, and the names were often descriptive (e.g., *whitefish* – a fish with white flesh). However, during the period of exploration, when an interest in global biodiversity began to emerge, these descriptive terms often became long and unwieldy as people sought to distinguish local species from similar relatives found elsewhere by adding additional descriptors (Winston, 2018). For example, in the 1600s, one type of *whitefish* was designated by the following Latin phrase that describes the shape and location of fins, the body length and width, and the mouth: *Gadus, dorso tripterygio, ore cirrato, longitudine ad latitudinem tripla, pinna ani prima ossiculorum trigiata* (Heard & Mlynarek, 2023, p. 2).

In the mid-1700s, Swedish biologist Carl Linnaeus systematized and formalized a new approach to scientific nomenclature that is still in use today: binomial nomenclature. This system uses a two-part Latin name to designate plants and animals: a single word for the genus followed by a single word for the species. Under this binomial nomenclature system, the unwieldy description for the fish mentioned above became simply *Merlangius merlangus*. A key feature of Linnaeus's system was that it decoupled naming from description. In other words, while the Linnaean binomials *may* be descriptive (e.g., referring to animal's shape), they do not *need* to be. According to Heard and Mlynarek (2023), this decoupling fundamentally changed the approach to naming species and opened the door to eponyms.

As explained by Winston (2018), binomial nomenclature caught on quickly and was almost universally adopted by Western science. Indeed, this system has been used to name hundreds of thousands of species. However, problems did arise, such as different groups of scientists assigning different scientific names to the same species. To prevent this, in the 19th century, a series of international codes emerged as a means of regulating nomenclature. The codes have gone through multiple revisions and versions, but they are still in use today. For instance, there is an International Code for Zoological Nomenclature and an International Code for Botanical Nomenclature, as well as three other codes for the naming of cultivated plants, bacteria and viruses. As Winston (2018) outlines, each code is accompanied by a protocol that scientists must follow to propose, to select, or on occasion, to replace scientific names. This is effectively a sort of terminology planning and management system for the biological sciences. As outlined by Sager (1990, p. 96), the following steps are required before a name can be fully accepted into a biological nomenclature:

1. The name must be constructed according to the rules.
2. Prior names must be considered since the oldest legitimate and properly constructed name claims precedence under the "Law of Priority".
3. The name must be accompanied by a complete description of the new species (i.e., list of attributes).
4. The name must be published in an established journal.

2.2 Eponyms in Scientific Names and Common Names

While binomial nomenclature is used to assign a unique Latin scientific name to a species, this species will likely have a common name in other languages too. For example, *Ursus maritimus* is the scientific name for the animal that is better known by most people by a common name, such as *polar bear* in English, *ours blanc* in French, or *Eisbär* in German. Eponyms can be integrated into the scientific name, the common name, or both.

In scientific names, the decoupling of naming and description paved the way for the use of eponyms. Linnaeus himself introduced a few eponymous terms, which he used mainly to pay homage to important scientists who had advanced the field, such as Matthias L'Obel and Jan Commelin, whose names are incorporated into the nomenclature through the genus names *Lobelia* and *Commelina* respectively (Heard & Mlynarek, 2023). However, while Linnaeus was restrained in his use of eponyms, their use has since proliferated, and not everyone who has been honoured was an eminent scientist (Poulin et al., 2022).

Meanwhile, a common name might be a (partial) translation of the Latin scientific name, and if that scientific name contains an eponym, this could be transferred to the common name. For example, the flowering plant *Lobelia erinus* is commonly referred to simply as *lobelia*. However, in other cases, the scientific name and the common name can be completely distinct, meaning that a scientific name that does not incorporate an eponym can be associated with a common name that does include an eponym. For instance, the bird *Nucifraga columbiana* is commonly known as *Clark's nutcracker* in English (named for William Clark who first spotted it during the Lewis and Clark Expedition), meaning that it has an eponymous common name, even though the scientific name contains no eponym.

Sometimes the scientific and common names may contain different eponyms, which happens in the case of *Ichterus parisorum*. This bird's scientific name honours the Paris brothers, who were French financiers in the early 1700s, while the bird's common name in English is *Scott's oriole*, named for Winfield Scott, a Confederate Army general during the U.S. Civil War.

3. Issues Identified With the Use of Eponyms

For several reasons, activists in both the scientific community and the general population have begun to advocate for changes to eponymous scientific and common names. Firstly, as noted above, eponyms are less transparent than descriptive names. Goldsmith (2023), among others, has pointed out that using descriptive common names rather than eponyms makes it easier for people to remember these names and understand some concrete details about the species. In turn, this is one way to improve public engagement, which is an essential step for future conservation efforts (AOS, 2023; Guedes et al., 2023).

Beyond transparency, social justice and EDI-oriented initiatives provide other excellent motivations for re-examining the use of eponyms. Indeed, within the grass roots of the scientific community, the practice of eponymous naming has come under scrutiny, mainly for two related reasons:

First, the *set* of people ‘honoured’ ... by eponymous names does not yet do a good job of representing the diversity of the human species. Second, some of the *individual* people chosen for eponymy are at least in hindsight regrettable choices. (Heard & Mlynarek, 2023, p. 6)

When considered globally, the set of eponymous terms demonstrate clear patterns: most refer to white men from the Global North (Heard & Mlynarek, 2023). For instance, Pillon (2021) investigated naming traditions for endemic plants of New Caledonia, an archipelago in the South Pacific that is recognized as a biodiverse region with many plants not found elsewhere. New Caledonia was colonized by and has been administered by France since 1853. Pillon’s (2021) study shows that 25% of the endemic taxa are named after people, and of these, 63% are named for French botanists. Swiss, German and British scientists are also honoured, particularly those who had travelled to the islands to collect specimens that were then taken back to their home countries for further study. Meanwhile just 7% of the species were named for New Caledonians.

Regarding gender, Pillon (2021) estimates that only 6% of the eponymous names of the plants in New Caledonia refer to women. Likewise, Figueiredo and Smith (2010) found that in the plant group *Aloe*, the number of species named for male scientists outnumber those named for women scientists by more than ten to one. Poulin et al. (2022, p. 7) make a similar observation about the naming of new species of parasites, noting

women researchers are under-represented among species named after eminent scientists. This is true across all parasite or host taxa, with the gender bias showing no evidence of improving over time in the past two decades. [...] The gender bias observed is certainly not owing to a shortage of excellent female taxonomists.

When it comes to the individuals whose names have been incorporated into eponymic species names, we find some very unworthy individuals, including Adolf Hitler (*Anophthalmus hitleri* or Hitler’s beetle), as well as George Hibbert (e.g. *Hibbertia hypericoides*), a wealthy patron of botany who was also a prominent slave-trader and slave-owner (Hammer & Thiele, 2021; Heard & Mlynarek, 2023). Likewise, the common names of species such as Couch’s kingbird are associated with colonizers who facilitated the displacement of Indigenous people from their homelands in the U.S. between 1830 and 1850 along a route now known as the Trail of Tears (Goldsmith, 2023).

Finally, some scientists use eponyms to make jokes, to dishonour rivals, or for other less-than-noble reasons (Heard & Mlynarek, 2023). This does not always sit well with other scientists, who feel that “The Earth’s biodiversity is part of a global heritage that should not be trivialized by association with any single human individual” (Guedes et al., 2023, p. 1159).

4. Approaches to Rectifying the Use of Eponyms

Activists in the scientific and broader community are calling for problematic names to be corrected, and for new naming guidelines to be developed. However, this movement is still in its early stages, and there is not yet consensus on the best ways of doing so. Overall, there appears to be goodwill to address the issues, but concerns that have been raised include the potential for introducing confusion into scientific exchanges, as well as the cost of updates (Ceríaco et al., 2023). Below we outline some emerging strategies for dealing with eponyms, as well as proposals for new naming techniques.

4.1 Updating Terminology Planning Processes

Firstly, as noted above, it is important to recognize that terminology planning is already a well-established activity in the biological sciences, with goals that include promoting terminological stability and precision. The previously mentioned International Committee on Taxonomy of Viruses (ICTV) is one such example, and this committee has developed guidelines for naming viruses. In the field of botany, the *International Code of Nomenclature for algae, fungi, and plants* – known informally as the *Shenzhen Code* – is the set of rules and recommendations that govern the scientific naming of all organisms traditionally treated as algae, fungi, or plants. Currently, the official rules enshrined in the Code do not permit renaming species that have a troubling or inappropriate name. Several botanists (e.g. Smith & Figueiredo, 2021) have therefore proposed additions and amendments to various articles in the *Code* that would allow names that are insulting, offensive or culturally inappropriate to be rejected in favour of more suitable names. Hammer and Thiele (2021, p. 1392) have gone even further by recommending that a new Permanent Nomenclature Committee be established – a Nomenclature Committee on Culturally Offensive or Inappropriate Names – to be elected by an International Botanical Congress in order to govern the application of the proposed articles.

Meanwhile, the American Ornithological Society (AOS) – a U.S.-based global network of scientists and bird lovers working together to advance the scientific study and conservation of birds – has also proposed some updates for terminology planning activities. In a press release on November 1, 2023, the AOS announced that it “will conduct an open, inclusive, and scientifically rigorous pilot program in 2024 to develop its new approach to English bird names in the U.S. and Canada” (AOS, 2023). As part of this pilot program, the AOS will establish a new committee to oversee the assignment of all English common names for species within the AOS’s jurisdiction. This committee will have a broad participation by including not only individuals with expertise in ornithology and taxonomy, but also those with expertise in communication and social sciences, as well as members of the general public.

Similarly, the Entomological Society of America’s (ESA) mission includes establishing effective communication between scientists, policy makers and the public. However, some eponymous names currently hinder effective communication. To address

problematic names, the ESA established a task force in 2021 and launched the Better Common Names Project, which has a terminology planning component with a five-step process for proposing and adopting better insect names (Lancette, 2021).

In Hawai'i, a cultural working group was established by locals in 2012 to advocate on behalf of Native Hawaiians (Gregg, 2021). Motivated by a desire to revitalize the Hawaiian language, the cultural group has since formed a nomenclature subcommittee, which brings together elders, cultural practitioners, scholars, resource managers and scientists with a view to giving Hawaiian names to plants and animals in the region. Under the auspices of terminology planning, which may be more formalized in some communities and less formalized in others, the following three general practices have emerged as approaches for dealing with eponyms: using more diverse and inclusive eponyms; re-instating former terms; and replacing eponyms with more transparent terms.

4.2 Using More Diverse and Inclusive Eponyms

While some activists argue that all existing eponyms should be replaced and no new eponyms should be permitted moving forward (e.g., Guedes et al., 2023), others see a complete ban on eponyms as a missed opportunity to empower in-country researchers to name their own biodiversity in a way that honours and celebrates local figures. Orr et al. (2023, p. 1168) even go so far as to note that the elimination of locally selected eponyms could sow discontent among these communities, noting that “One might even consider a ban on eponyms itself colonialistic without consideration or due compensation for the hundreds of years that colonizing countries pursued such goals unchecked.”

Some communities are instead electing to diversify the profile of those for whom species are named (Antonelli et al., 2023). According to Pillon (2021), not only does the (re-)naming of species provide an opportunity to acknowledge more broadly the diversity of individuals who have contributed to our understanding of the natural world, it may also aid ecological conservation. With this in mind, Pillon (2021) recommends that species should be named with a view toward how these names will be perceived by the local communities involved.

4.3 Re-Instating Former Terms

Numerous species were claimed as new discoveries by explorers and colonial settlers, even though these were already known to the Indigenous communities. Some scientists advocate for restoring Indigenous names in taxonomy, noting that such names can often be knowledge conduits, conveying information about history, place, and belonging. For instance, Gillman & Wright (2020) argue that although the “Law of Priority” (see section 2.1) is a fundamental element of the binomial nomenclature system, the chronological precedence of Indigenous names has no standing or priority under current taxonomic codes, even though these names often convey in-depth knowledge relating to the species. Gillman & Wright (2020) propose taxonomic

rule changes to promote retrospective name changes that establish, on the basis of precedence, pre-existing Indigenous names for species where possible. As a key first step, Gillman & Wright (2020) call for a general debate on the merits of such an approach, emphasizing that the voices of Indigenous scientists and their communities must be central to this debate.

As for common names, members of a cultural group in Hawai'i conduct research to identify the original names for local species that were lost when the Hawaiian language was repressed during the long period of Western encroachment (Gregg, 2021). In 2017, a member researched *Tristram's storm petrel* (named after a British scientist) and uncovered the traditional name for this bird – 'ākihike'ehi'ale – in a Hawaiian scholarly text dating from 1860. This traditional name has now been reinstated.

This approach may soon happen at scale as Indigenous communities around the world continue reclaiming their languages, so it is worth discussing this approach in Terminology, perhaps using *retro-terminologization* to describe the practice. The notions *terminologization* and *de-terminologization* are already used to describe cases where general language words move into specialized language, and vice versa (Infoterm, 2005). *Retro-terminologization* could describe cases where an original term that had been set aside is later restored to use.

4.4 Replacing Eponyms With Transparent Terms

Because eponyms are not transparent, some groups are replacing them with descriptive names. For instance, in 2020, the AOS chose to replace the common name *McGown's longspur* (originally named for John P. McGown, a naturalist who later joined the Confederate Army during the U.S. Civil War) with the more transparent term *thick-billed longspur* (AOS, 2023).

Just as pre-existing Indigenous names can be reinstated, so too can Indigenous knowledge be used to form new names (Veale et al., 2019). Hágaster & Wrazildo (2020) took this approach to find a name for a species of orchid discovered in Venezuela. The Indigenous community proposed the name *Epidendrum katarun-yariku* because, in the Pemón Arekuna language, *katarun* means high and *yariku* means flower – a nod to the fact that this species only grows on the high cliffs. Meanwhile, Gregg (2021) reports that the cultural group in Hawaii has adopted a similar strategy, working with elders to create new common names for regional birds. In this way, *Christmas shearwater* has been replaced by 'ao'ū, an onomatopoeic name that evokes the bird's nocturnal flight call, while the *Bonin petrel* has been given the new name *nunulu*, which is a Hawaiian word meaning warbling.

5. Conclusions and Future Directions

Eponyms can inflict harm and can hinder precise communication through non-transparency. The grassroots movement to decolonize species' names in the biological sciences and to make them more diverse and inclusive is gaining traction, but there

is not yet consensus and major changes will take time. Currently, the movement to replace harmful eponyms is coming from within the biological sciences, where bodies are already in place to guide terminology planning. However, scientific names are often accompanied by common names, which may also need to be replaced. Several projects are underway that welcome input from specialists in different areas (including language) as well as from the community more broadly. With their expertise in term formation, terminologists could contribute to the discussions.

This article focuses on biological sciences, but eponym use may need to be re-examined elsewhere (e.g., medicine). Moreover, eponyms are just one type of problematic term. Smith and Figueiredo (2021) identify practices leading to biased terms and have submitted a proposal to permanently and retroactively eliminate racist scientific names from the nomenclature. Others challenge ableist and sexist terms (Parsley, 2020; Rummy & Rummy, 2021). There is still work to do at the intersection of Terminology and EDI, but as Cheng et al. (2023) note, identifying harmful terms may not be easy since meaning changes across languages, cultures, contexts and time. Terminologists, who have experience working in multilingual and multicultural contexts along with methodological expertise for identifying conceptual characteristics and proposing transparent terms, can help to make specialized communication more inclusive.

References

- American Ornithological Society (AOS) (2023). AOS Will Change the English Names of Bird Species Named After People. <https://americanornithology.org/american-ornithological-society-will-change-the-english-names-of-bird-species-named-after-people/>
- Antonelli, A., Farooq, H., Colli-Silva, M., Araújo, J., Freitas, A., Gardner, E., Grace, O., Gu, S., Marline, L., Nesbitt, M., Niskanen, T., Onana, J., Pérez-Escobar, O., Taylor, C., & Knapp, S. (2023). People-inspired names remain valuable. *Nature Ecology & Evolution*, 7, 1161–1162. <https://doi.org/10.1038/s41559-023-02108-7>
- Ceríaco, L., Aesch, E., Ahyong, S., Ballerio, A., Bouchard, P., Bourgoïn, T., Dmitriev, D., Evenhuis, N., Grygier, M., Harvey, M., Kottelat, M., Kluge, N., Krell, F., Kojima, J., Kullander, S., Lucinda, P., Lyal, C., Pyle, R., Rheindt, F., Scioscia, C., Welter-Schultes, F., Whitmore, D., Yanega, D., Zhang, Z., Zhou, H., & Pape, T. (2023). Renaming taxa on ethical grounds threatens nomenclatural stability and scientific communication. *Zoological Journal of the Linnean Society*, 197, 283–286. <https://doi.org/10.1093/zoolinnean/zlac107>
- Cheng, S., Gaynor, K., Moore, A., Darragh, K., Estien, C., Hammond, J., Lawrence, C., Mills, K., Bais, M., Ignace, D., Khadempour, L., McCary, M., Rice, M., Tumber-Dávila, S., & Smith, J. (2023). Championing inclusive terminology in ecology and evolution. *Trends in Ecology & Evolution*, 38(5), 381–384.
- Enserink, M. (2013). International group settles on new name for coronavirus. *Science*. doi: 10.1126/article.24291
- Figueiredo, E., & Smith, G. (2010). What's in a name. *Bradleya*, 28, 79–102.

Gillman, L., & Wright, S. (2020). Restoring indigenous names in taxonomy. *Communications Biology*, 3, 609. <https://doi.org/10.1038/s42003-020-01344-y>

Goldsmith, S. (2023). The Movement to Rename Species. *Nature Conservancy Magazine*, Fall. <https://www.nature.org/en-us/magazine/magazine-articles/new-common-names/>

Gorbalenya, A., Baker, S., Baric, R., de Groot, R., Drosten, C., Gulyaeva, A., Haagmans, B., Lauber, C., Leontovich, A., Neuman, B., Penzar, D., Perlman, S., Poon, L., Samborskiy, D., Sidorov, I., Sola, I., & Ziebuhr, J. (2020). The Species Severe Acute Respiratory Syndrome-related Coronavirus. *Nature Microbiology*, 5, 536–44. <https://doi.org/10.1038/s41564-020-0695-z>

Gregg, J. (2021) Decolonizing seabirds. *Hakai Magazine*. <https://hakaimagazine.com/news/decolonizing-seabirds/>

Guedes, P., Alves-Martins, F., Martínez Arribas, J., Chatterjee, S., Santos, A., Lewin, A., Bako, L., Webala, P., Correla, R., Rocha, R., & Ladle, R. (2023). Eponyms have no place in 21st-century biological nomenclature. *Nature Ecology and Evolution*, 7, 1157–1160. <https://doi.org/10.1038/s41559-023-02022-y>

Hágsater, E., & Wrazidlo, M. (2020). *Epidendrum katarun-yariku* (Orchidaceae), a new species of the *Schistochilum* group from the tepuis of the Guiana Highlands in South America. *Phytotaxa*, 472(1), 33–40.

Hammer, T., & Thiele, K. (2021). Proposals to amend Articles 51 and 56 and Division III, to allow the rejection of culturally offensive and inappropriate names. *Taxon*, 70(6), 1392–1394. <https://doi.org/10.1002/tax.12620>

Heard, S., & Mlynarek, J. (2023). Naming the menagerie. *Proceedings of the Royal Society B*, 290, 20231970.

Infoterm. (2005). *Guidelines for Terminology Policies*. UNESCO.

Lancette, J. (2021). Breaking Barriers in Entomology: The Better Common Names Project. *American Entomologist*, Winter, 10–11. <https://doi.org/10.1093/ae/tmab061>

Lee, E., & Johnstone, M. (2021). Resisting the politics of the pandemic and racism to foster humanity. *Qualitative Social Work*, 20(1-2), 225–232. doi: 10.1177/1473325020973313

Orr, M., Hughes, A., Torres Carvajal, O., Ferrari, R., Luo, A., Rajaei, H., Ron, S., Warrit, N., Zamani, A., Zhang, Y., & Zhu, C. (2023). Inclusive and productive ways forward needed for species-naming conventions. *Nature Ecology & Evolution*, 7, 1168–1169. <https://doi.org/10.1038/s41559-023-02103-y>

Parsley, K. (2020). Plant awareness disparity: A case for renaming plant blindness. *Plants, People, Planet*, 2(6), 598–601. <https://doi.org/10.1002/ppp3.10153>

Pillon, Y. (2021). The inequity of species names: The flora of New Caledonia as a case study. *Biological Conservation*, 253, 108934. <https://doi.org/10.1016/j.biocon.2020.108934>

Poulin, R., McDougall, C., & Presswell, B. (2022). What's in a name? Taxonomic and gender

biases in the etymology of new species names. *Proceedings of the Royal Society B*, 289: 20212708. <https://doi.org/10.1098/rspb.2021.2708>

Rummy, P., & Rummy, J. (2021). Recontextualizing the style of naming in nomenclature. *Humanities & Social Sciences Communications*, 8, 283. <https://doi.org/10.1057/s41599-021-00975-8>

Sager, J. (1990). *A Practical Course in Terminology Processing*. John Benjamins.

Smith, G., & Figueiredo, E. (2021). (126) Proposal to add a new Article 61.6 to permanently and retroactively eliminate epithets with the root caf[e]r- or caff[e]r- from the nomenclature of algae, fungi and plants. *Taxon*, 70(6), 1395–1396. <https://doi.org/10.1002/tax.12622>

Veale, A., de Lange, P., Buckley, T., Cracknell, M., Hohaia, H., Parry, K., Raharaha-Nehemia, K., Reihana, K., Seldon, D., Tawiri, K., & Walker, L. (2019). Using te reo Māori and ta re Moriri in taxonomy. *New Zealand Journal of Ecology*, 43(3), 3388. <https://dx.doi.org/10.20417/nzj ecol.43.30>

Winston, J. (2018). Twenty-first Century Biological Nomenclature. *Integrative and Comparative Biology*, 58(6), 1122–1131. <https://doi.org/10.1093/icb/icy060>

Acknowledgements

Partly funded by the Faculté des lettres et des sciences humaines and the Presses de l'Université Laval.

Contact information

Lynne Bowker
Université Laval
lynne.bowker.1@ulaval.ca

Theresa Kruse, Ulrich Heid, and Barbara Schmidt-Thieme

MATHEMATICS STUDENTS AS LEXICOGRAPHERS

Learning Domain Concepts and Their Relations by Designing Dictionary Articles and Concept Maps

Abstract We present a study which was carried out with teacher students of mathematics. They were asked to create either dictionary articles or concept maps for terms from an introductory lecture in their first semester. Based on the students' submissions, we investigate whether there is a difference in the learning outcomes between the two tasks and also whether the technical means used to solve these tasks influence the students' engagement in the tasks, i.e., whether they chose digital tools or handwriting to complete them. The analysis presented here is based on a first annotated subset of our data and provides preliminary results on our research questions. We show that digital tools seem to be more appropriate to motivate a deeper exploration of the domain. In addition, our analysis suggests that dictionary articles and concept maps motivate different cognitive approaches to the domain, depending on whether the focus is on the concepts themselves or on the possible relations between them.

Keywords learning; mathematics; concept maps; dictionary articles

1. Introduction

In introductory lectures in universities, mathematics tends to be presented as a sequence of definitions, theorems, and proofs, while in school it is more often presented as a collection of algorithms. Our goal is to make first-year students aware of the fact that mathematics is actually a web of concepts that are connected by different kinds of relations. We hypothesize that this can be achieved if students have to describe concepts and the relations between them in the way in which they are presented in dictionaries and concept maps. Thus, we gave the students the task of writing dictionary articles and creating concept maps.

In this paper, we focus on three research questions:

(RQ1) Is there a relation between mathematical knowledge and engagement¹ in solving dictionary or concept maps tasks, respectively? And if so, how strong is it?

(RQ2) Are there differences in the engagement between dictionary and concept map tasks? And if so, what are they?

(RQ3) Does the engagement in dictionary and concept map tasks depend on the mean used to solve them? And if so, to what extent?

¹ By engagement we mean here the intensity with which the students have worked on the task, measured in the number of mentioned concepts and relations.

We present preliminary answers to these questions based on a subset of the data we collected: We look at the first tasks handed in by the students. We present the setting in which we conducted our study in Section 2, while the actual analysis is given in Section 3. Section 4 summarizes the results and our learnings. Limitations and a final discussion are given in Section 5.

2. Experimental Setting

The participants of our study are students in the first year of their studies to become mathematics teachers. The course was carried out in the winter term of 2023/24. At the beginning of the semester, 167 students enrolled in the course, out of which 44 were assigned to create concept maps, 50 to create dictionary articles and 73 did other terminology-related tasks. The allocation of the students to the three groups was not randomized but depended on the time slot they chose for their tutorial session. Thus, external factors such as time slots occupied by courses in their second field of study may have influenced the distribution. This also explains the different group sizes.

Each week, the students were given a choice task and a compulsory task. The choice tasks were the same for all students and comparable to classical mathematical tasks at university, i.e., calculating or proving things. In the following evaluation, we use the choice task for estimating the general mathematical abilities. The compulsory tasks differed between the students. The first group always had the task of writing dictionary articles for two given concepts, the second group always had the task of creating a concept map containing the two same concepts, and the third group was a control group that was given either the dictionary task, the concept map task, or another terminology task. There were ten of these tasks over the semester. On average, 110 people submitted their solutions each week. This number is lower than the 167 persons who originally signed up for the course because of dropouts and course repeaters. Over the semester, we collected a total of 470 submissions for the dictionary article tasks, 329 submissions for the concept maps, and 303 submissions for the other tasks. At the beginning of the semester, the students were given a short introduction to the production of dictionary entries and concept maps, covering both, structural and infrastructural aspects, such as recommended software. They were also informed about the fact that we conduct a study accompanying the course.

3. Analysis of Students' Submissions

In this section, we present a first analysis of the students' submissions from which we generate hypotheses for answering the research questions which have to be evaluated on the whole data set.

4. Dictionary Articles

The purpose of the dictionary articles can be described in the framework of the function theory (Fuentes-Olivera & Tarp, 2014). In our case, the users are the same group as the lexicographers as they are writing the dictionary entries for themselves,

for example for using them as a learning tool at a later point. Further users are the lecturers as they evaluate the results, but as this is part of the meta-setting, we do not count the lecturers as users in the narrower sense. The students are writing the entries in (one of) their first language(s). Concerning the lexicographic aspects, they are lay people, and they are semi-experts in the domain of the entries (Bowker, 2003). The usage situation is a cognitive one as its main purpose is to gain or deepen knowledge about the domain. The situations could be both, sporadic and systematic, as there could be sporadic situations where a user wants to look up a certain meaning, while the overall interest is a systematic one, i.e., learning and understanding the concepts of the given domain.

The following data is based on the first dictionary task in which the students were asked to create articles for the terms *Wertemenge* (i.e., *range of a function*) and *Definitionsmenge* (i.e., *domain of a function*). 40 persons handed in their solutions.

First, we look at the items they use in their article structures and give the percentage of students who used a given item in their article structure in parentheses: definition (100%), synonyms (82.5%), hypernyms (82.5%), hyponyms (77.5%), associated concepts (77.5%), collocations (70%), Part of Speech labels (PoS) (5%). Two further item types, namely ‘further information’ and ‘variant’ are used by one person each. Besides PoS, all the information provided belongs to the comment on semantics and is mostly identical to typical items in specialized dictionaries (Humbley, 2017).

We also annotate the technical mean used by the students to create the articles and find five categories: editor (42.5%), XML (27.5%), tablet (20%), hand-written (5%), table (5%). With the category “editor” we refer to dictionary entries written with a text editing program. For highlighting the structural indicators, different strategies like bold typeface are used. In the table category, the entries are provided in a table. The XML entries have been created with Lexonomy (Měchura, 2017) and exported. The category ‘tablet’ comprises entries created using a tablet, which is a variant of the use of handwriting but provides a wider range of possibilities for a more flexible editing due to its digital nature. This flexibility distinguishes them from the submissions in the category hand-written which are scans or photos of hand-written articles on paper.

All students use a panel view for presenting their data (Koplenig & Müller-Spitzer, 2014). Single students also create a kind of access structure, e.g., by an index, but this was not part of the task they got.

5. Concept Maps

We analogously evaluate the first task for the group who created the concept maps for the same two terms, *Wertemenge* and *Definitionsmenge*. It was handed in by only 25 persons. For the annotation of the categories, we use a deductive approach based on a category system for relations between mathematical terms by Kruse et al. (2023). The most common categories in the task we analyze are (with the percentage of

students who used them given in parentheses): synonymy (92%), hypo-/hyperonymy (84%), part-of-relation (80%), mapping (32%), causality (32%), analogy (8%), property (4%), medium (4%). In addition, there are connection lines in the maps without a named relation (44%) and others that could not be mapped to one of the categories of the scheme but have been only used by single students (24%).

Analogously to the dictionary task, we also evaluate the means the concept maps were created with. We find the following categories analogously to the dictionary entries: hand-written (24%), with a tablet (44%), or with a digital tool for creating concept maps (32%).

6. Results of Quantitative Evaluations and Learnings

For the quantitative analysis of the students' submissions, we carry out t-tests for comparing different subgroups and we calculate correlation coefficients for finding relations between different aspects. As a reference, we use the points achieved in the choice task which was handed in with the dictionary or concept map tasks, respectively ('points'). Furthermore, we consider the number of different items in the article structure for the dictionary task which we regard as equivalent to the number of different relations in the concept map task ('relations'). The third number we consider is the total number of items attributed to the structural indicators ('items').

First, we look at correlations between the different values. For the concept maps, we find no significant pairwise correlations (see Table 1). The reason for that could be that we only have a small amount of data in that first task, with 25 participants. However, the correlations for the dictionary task are significant and we find in all three cases positive correlations with medium to large effect (Cohen, 1988). It would be interesting to see if the numbers for the concept maps give similar results once calculated for the bigger data set. For now, we can conclude that students who acquire more points in other mathematical tasks tend to create bigger dictionary articles. Nevertheless, this correlation might also be caused by other factors not considered in our survey.

Table 1: Results of Pearson correlations between items, relations and points achieved

	Dictionary articles (N=40)	Concept maps (N=25)
Items and relations	0.8371 (p<0.001)	0.3492 (p=0.0871)
Items and points	0.4768 (p<0.01)	-0.1188 (p=0.5716)
Relations and points	0.5989 (p<0.001)	-0.3629 (p=0.0746)

Thus, we investigate if there are significant differences between both groups. The results in Table 2 show that there is no significant difference in the achieved points (out of 20) between the two groups. However, in the dictionary articles, significantly more different relations are used than in the concept maps. The reason for that could

be that both formats favor different approaches in the creation, as someone who is writing a dictionary article might be more focused on adding more items to the article structure, while someone creating a concept map is more focused on adding more concepts rather than finding more different relations. This could be also reflected by the fact that there is no significant difference in the total number of items given between both groups.

Table 2: Arithmetic means and t-test results

	Dictionary articles (N=40)	Concept maps (N=25)	Significance
Points	14.63	15.98	n.s.
Relations	5.20	4.04	p<0.01
Items	19.5	17.2	n.s.

The third aspect we evaluate is the mean which the students use to create their dictionary articles or concept maps, respectively. Therefore, we divide the categories into two groups, namely digital and hand-written. The idea behind this is mainly that students who are typing are more flexible in their editing than when doing it by hand where they are more bound to an established structure. Nevertheless, we have to keep in mind that our analyses are only based on the final submissions and we do not have access to the students' work processes where different draft versions might have been created.

Concerning the amount of points reached in the choice tasks, we find no significant differences in the preference for one or the other mean. Significant differences are only found for the dictionary entries where more relations and items are added in the digital tools, while the results are not significant for the concept maps. All in all, we may come to the hypothesis that the digital format motivates students to add more information in their dictionary entries and concept maps, which, in turn, might lead to a deeper engagement with the topic. Thus, we conclude that students should be motivated and carefully instructed to use digital tools for creating dictionary entries or concept maps if the goal is to motivate them for looking deeper into the domain.

Table 3: Results of the t-tests for digital and hand-written submissions. For each item, we give the arithmetic mean as well as the significance

	Digital	Written	Significance
Dictionary articles: points	16.0	10.5	n.s.
Dictionary articles: relations	5.97	2.90	p<0.01
Dictionary articles: items	22.6	10.2	p<0.01
Concept maps: points	17.56	15.24	n.s.
Concept maps: relations	3.88	4.12	n.s.
Concept maps: items	21.0	15.41	n.s.

7. Discussion

Based on our quantitative evaluations we can provide preliminary indications for answering our research questions. However, further investigation with a bigger data set is necessary to evaluate if these tendencies are robust.

While we cannot give a final answer to (RQ1) for the concept map tasks, we can say for the dictionary task that there is a tendency towards a connection between higher mathematical competence and more engagement in solving the task. However, we cannot say anything about the direction of this relation based on our survey setting. Also, external motivation factors could influence both aspects.

Concerning (RQ2), both kinds of tasks seem to yield similar results concerning the engagement, though the focus is a bit different. If the didactical focus is on finding more different kinds of relations between terms, we recommend choosing dictionary tasks.

With our results for (RQ3), we recommend letting students create both, concept maps and dictionary entries, with digital tools to gain a deeper engagement. The reason for that might be that the tools enable changes in the structure and, along with it, rethinking about the task more easily.

In terms of future work, we plan to expand our evaluation to the whole student data to see if our hypotheses still hold. Additionally, it would be interesting to investigate whether the results are transferable to other domains of the exact sciences as well or mainly concern the didactics of mathematics.

References

- Bowker, L. (2003). Specialized lexicography and specialized dictionaries. In P. G. J. van Sterkenburg (Ed.), *A practical guide to lexicography*. (pp. 154–164). John Benjamins.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2. Aufl.). L. Erlbaum Associates.
- Fuertes-Olivera, P. A., & Tarp, S. (2014). *Theory and Practice of Specialised Online Dictionaries. Lexicography versus Terminology*. De Gruyter.
- Humbley, J. (2017). Specialised dictionaries. In P. A. Fuertes-Olivera (Ed.), *The Routledge Handbook of Lexicography*. Taylor & Francis Group. <http://ebookcentral.proquest.com/lib/ubhildesheim-ebooks/detail.action?docID=5088188>
- Koplenig, A., & Müller-Spitzer, C. (2014). Questions of design. In C. Müller-Spitzer (Ed.), *Using online dictionaries* (pp. 189–204). De Gruyter.
- Kruse, T., Heid, U., & Girnat, B. (2023). Using lexicography for learning mathematics. In M. Medved', M. Měchura, C. Tiberius, I. Kosem, J. Kallas, M. Jakubiček, & S. Krek (Eds.), *Electronic lexicography in the 21st century (eLex 2023): Invisible Lexicography. Proceedings of*

the eLex 2023 conference (pp. 466–475). <https://elex.link/elex2023/wp-content/uploads/97.pdf>

Měchura, M. (2017). Introducing Lexonomy: An open-source dictionary writing and publishing system. In I. Kosem, C. Tiberius, M. Jakubíček, J. Kallas, S. Krek, & V. Baisa (Eds.), *Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference* (pp. 662–679).

Contact information

Theresa Kruse

University of Hildesheim
theresa.kruse@fau.de

Ulrich Heid

University of Hildesheim
heidul@uni-hildesheim.de

Barbara Schmidt-Thieme

University of Hildesheim
bschmidt-thieme@imai.uni-hildesheim.de

David Lindemann

TEACHING TERMINOLOGY THROUGH WIKIBASE AND WIKIDATA

Abstract This software demonstration presents approaches to employ Wikibase in a university course on Terminology, and results of terminology projects lead by students. Wikibase, an extension of MediaWiki, is the software that underlies Wikidata, a very large crowdsourced queriable knowledge graph. We use an own Wikibase instance for cloud-based collaborative student projects on Terminology in Basque, a European minority language spoken in Spain and France, and co-official in some regions of Spain. We show that the course work with and about this software covers a great deal of what a course in Terminology for students in the translator and interpreters training should cover; it embraces relevant facets of knowledge engineering, and the acquisition of the features of a graph database and its user interface. The datasets created in student projects and partially shared through Wikidata are examples for a low-barrier crowdsourcing terminology workflow, eventually useful in other contexts as well.

Keywords terminology; teaching; knowledge graphs; Wikibase; Wikidata

1. Introduction

Understanding structures of terminological databases is regarded a fundamental skill in Translation Studies (see, for example, Cabré, 2000), and knowing to implement workflows for the creation of digital terminological resources might be necessary for developing a terminological Bachelor Final Project, while it certainly is indispensable for Master Final Projects in Terminology.¹ At the same time, we believe that challenging students with problems of practical terminology work motivates them to dig into multiple theoretical aspects of Terminology like conceptual relations, word formation, neology, and linguistic purism, and into the representation of knowledge in formal models in general. Also, in the attempt to write about their own endeavours, students need to acquire adequate meta-terminology, which we believe they will retain better when it has immediate application in an own project.

The Terminology course we offer for Translation Studies Bachelor students is introduced by providing fundamental concepts and problems of Basque and general Terminology Science,² but then deals with basic lexicographical and terminological resource structures, including formal representations of lexical-semantic relations, showing examples for the different kinds of lemma-focused

¹ See, for example, the module description of the Online Master in Terminology at UPF, Barcelona: <https://www.upf.edu/web/terminologiaonline/memoria-de-master>.

² We do not have space here for listing a comprehensive bibliography. Basque Terminology Science is heavily inspired by Catalan publications by Cabré et al.; some of them have been translated.

and concept-focused³ terminological resources: LSP dictionaries, thesauruses, taxonomies, term banks and wordnets, in human-readable distribution formats, and as structured datasets in machine-readable representation formats. It is shown how the models applied in structured datasets find their counterparts in the layout given to the representation shown to the human end user. This relation is looked at from two perspectives: How a resource entry structure is translated into entry layout features, and how layout features could be used to define a schema when trying to structure a resource entry text. In fact, the latter is what students will be asked to do in their own projects: To look out for an existing resource containing Basque terminology for the domain of their choice, and to work on that, with the aim of integrating the Basque terms, and, eventually, their descriptions, in the Wikidata Knowledge Graph. They receive advice for defining an appropriate workflow. We also offer to employ an own script that extracts concepts appearing in a custom set of Wikipedia articles. This simple approach has turned out to be valuable for a range of projects.⁴ Alternatively, students might choose existing resources that contain terms in other languages as starting point; here, the mission is to define Basque equivalents for the described concepts, by examining specialized text written in Basque, or Basque terminological resources formerly not connected to the object of study.

All students know and heavily use Wikipedia in their homework and for documentation in their translation tasks, and practical translation courses at UPV/EHU that involve the Basque language address Wikipedia content creation by translation (most often from Spanish to Basque), which forces them to get along with requirements and editing procedures given on that platform,⁵ produces the satisfaction of having contributed to a globally known resource, and lowers the barrier for subsequent contributions.⁶ This is one strong reason for also looking at Wikidata (Vrandečić, 2012), Wikipedia's sister project at the Wikimedia Foundation, which started more than a decade ago as the structured, machine-readable counterpart of the human-centered and only roughly structured Wikipedia, but soon included also content from other sources as well, and which is now regarded one of the largest free knowledge graphs on the web, with significant conceptual coverage of important domains of knowledge (see e.g., Waagmeester et al., 2020). At the same time, the set of lexical-semantic relations represented in the Wikidata Knowledge Graph include all those found in the other lexical-conceptual resources presented, and Wikidata entities very often also include "external id" statements

³ We use the term *lemma-focused* for resources such as dictionaries, which use a lemma list as macrostructure, in opposition to concept-focused resources, which model concepts as nodes that relate to each other, like in a hierarchized thesaurus or a knowledge graph, and that are annotated with equivalent terms in one or more languages, i.e., the terms that are used in texts when referring to the concept. When comparing viewpoints and workflows in specialized lexicography, on one side, and terminology, on the other, we follow Costa (2013).

⁴ The script returns a list of Wikidata entities extracted from the blue-coloured hyperlinks in Wikipedia article text. After deleting concepts irrelevant to the project from that list, these are cloned on the local Wikibase instance, where they can be enriched with missing Basque equivalents, and other annotations.

⁵ Translators often have to deal with cloud-based software platforms designed by the client, and clients expect them to be fast and flexible in that sense. This aspect, in our opinion, is still not sufficiently acknowledged in Translation Studies curricula.

⁶ The author wants to thank Wikimedia Basque Country for their support in translation courses, providing a dedicated training library, and delivering tutorial seminars.

to those other resources on entity level.⁷ This and the fact that editing Wikidata requires not more than a Wikimedia user account makes it interesting to aim at enriching Wikidata as a student project.

As for the basic meta-terminology to use in the course, students are introduced to terms used in the description of term banks following Fontenelle & Rummel (2014) and of relevant representation formats following Reineke & Romary (2019), that is at its core, to describe *terms* as lexical units of a language that occur in specialized text, and that denote *concepts* as part of a specialized domain; if seen the other way round, that is, looking at a terminological resource, terms appear as lexicalizations or *labels* of concepts, which are connected to each other through typed relations. This is the way concepts are modelled on a Wikibase.⁸

2. Student Projects

In most cases, the sources chosen by the students will be unstructured data, i.e., a PDF publication, such as one of the specialized Basque dictionaries of the UZEI series,⁹ or, for example, a list of recipes and a glossary listed on a Basque cooking website. On the other side, if the chosen source is an established structured terminological resource, that resource will most probably be well covered with concept labels and textual definitions in English and other major languages, while the coverage of Basque equivalents to these will present gaps. The students' aim, accordingly, will be to fill these with data obtained in their own terminological project.

All source datasets are converted to Wikibase format, that is, concepts are represented as Wikibase *items*, i.e., as nodes in a graph, annotated with preferred and alternative labels and definitions (in multiple languages), and with conceptual and other relations as edges – the Wikibase *properties*. If the source is Wikidata, the entity subset to work on is cloned on the local Wikibase. This conversion is done by the teacher and made transparent to the involved students.

Students are familiarized with the Wikibase online interface, and with the Open Refine software application,¹⁰ which they install locally. The first step is to try to match the concepts of the students' concept schemes to Wikidata, using the application. For example, if the source to work with is a Basque-Spanish specialized dictionary, matching concepts can be looked for on Wikidata using the Basque and the Spanish term in a label-based search, eventually restricted to instances of a class (e.g., “instance of color“, for color names). After that, regardless of whether projects have started with an external resource uploaded to Wikibase, or with a subset of Wikidata, all

⁷ For example, alignments to Princeton WordNet are represented using the property <http://www.wikidata.org/entity/P8814>, and those to the EuroVoc thesaurus using <http://www.wikidata.org/entity/P5437>.

⁸ In addition, on a Wikibase, terms may be represented as *lexemes* (lexical entries), which allows for a richer linguistic or lexicographic description, following Ontolex-Lemon. See <https://eneoli.wikibase.cloud> for a recent example of describing terms as Wikibase lexemes.

⁹ The UZEI dictionaries are available as PDF at <https://www.euskadi.eus/hiztegiak-banku-terminologikoak-eta-entziklopediak/web01-a2eutres/eu/#9588>.

¹⁰ See <https://openrefine.org>.

students are at a similar point: They have a set of Wikibase entities, some or all of them aligned to Wikidata entities, and the labels, definitions, and, eventually, the relations contained as statements in Wikidata. Now, missing labels, descriptions, and relations can be added or updated. Basque concept lexicalizations proposed by students are to be furnished with usage attestations, which are referenced as Wikibase *reference* using an external URL or a Wikibase entity describing the cited resource. Concept relations to be worked on are most typically “instance of”, “kind of” a.k.a. “subclass of”, and “part of”,¹¹ but use cases may involve others (such as “used by”¹² for a scheme of social media concepts, or “has part” for a cooking concept scheme).¹³

Manual and batch edits to Wikibase entity data are tracked in the reversible Wikibase edit history, and collaborative work may involve the use of entity discussion pages and/or user talk pages. Furthermore, students use self-created Wikitext pages to describe their project; the teacher supports the process providing SPARQL queries tailored for each use case. Most typically, such query will list existing (Wikidata) item labels and descriptions, and their counterparts as they are described on the local Wikibase in that point in time. After a collaborative review process, project results can be uploaded to Wikidata. The uploaded datasets in all cases contain Basque concept lexicalizations and descriptions, and in some, also concept relations and entities to be created as new on Wikidata.

2.1 Example Project on Climate Change

The documentation of student projects carried out so far are accessible from the home page of our Wikibase instance, in Basque language.¹⁴ As example for a student project, we present here a recent piece of work presented as Bachelor Final Project in Basque, and briefly described in English:¹⁵ Paula Garay’s concept scheme on Climate Change terminology. The workflow can be described as follows:

1. Relevant authority sources are identified, in this case, glossaries released by government authorities, international organizations, and NGOs.¹⁶
2. The English terms described in these resources are uploaded to Wikibase as English labels for newly created concept (“Q”) items, together with English definitions given in these, referencing each term and each concept definition to its source. Since referencing labels in Wikibase is not possible, the terms and the definitions are represented as separate statements as well;¹⁷ this allows further descriptions of the lexical item considered an equivalent term in a language, using Wikibase qualifiers and references.¹⁸

¹¹ That is, the three fundamental concept relations according to ISO 25964-2, a norm dealing with the interoperability of terminological resources; see <https://www.iso.org/standard/53658.html>.

¹² On the own Wikibase, represented as <https://eusterm.wikibase.cloud/entity/P50>.

¹³ See the ingredients of a lasagne on Wikidata at <https://www.wikidata.org/wiki/Q20034#P527>.

¹⁴ See <https://eusterm.wikibase.cloud>.

¹⁵ See https://eusterm.wikibase.cloud/wiki/GRAL_Paula_Garay.

¹⁶ See exact references on the project page.

¹⁷ For an example, see <https://eusterm.wikibase.cloud/wiki/Item:Q6981#P94>.

¹⁸ See <https://www.mediawiki.org/wiki/Wikibase/DataModel/Primer>.

3. In Open Refine, the English terms are used for reconciliation against Wikidata entities, and the Wikidata entity ID of matching candidates are added to the Wikibase entities.
4. Through a federated SPARQL query, a table is produced, which contains the source term and definition, and the definition of the matching candidate on Wikidata. In this step, the student validates the alignment of Wikibase and Wikidata entity, and updates it if, looking at the English concept description found on Wikidata, the alignment is not regarded an exact match.
5. Wikibase entities that have been linked to one and the same Wikidata entity are merged. These are those entities that describe terms stemming from different sources but that describe the same concept. Consequently, these are then described on the same Wikibase entity page.¹⁹
6. The resulting alignment contains 376 concepts, from which 200 had a Basque label on Wikidata (62 had Basque definitions). These have been imported to Wikibase, together with labels in four other languages that the student understands (German, Spanish, Catalan, and French), the links to access Wikipedia articles describing the concept in these languages, and statements (property-value pairs) for the following properties, if present on Wikidata: *instance of*, *part of*, *subclass of*.
7. The concept labels imported from Wikidata are reviewed. In this context, the function of Wikidata labels is discussed, which is not fully aligned to terminology criteria; more precisely, “alternative labels” on Wikidata are often not synonyms of the “preferred label”, that is, they are not interchangeable terms. Wikidata labels are chosen for the retrieval of items through text-based entity searches; “alternative” labels are, for example, found to be hyperonyms of the “preferred” label (Martín-Chozas et al., 2020).
8. Basque term gaps that the student has found candidates for, or term labels stemming from Wikidata found valid, are annotated with an authoritative source, and a context snippet. This is then represented as a statement to the Wikibase entity, and can be used as source for an upload to Wikidata, where the Basque term will henceforth serve as preferred or alternative concept label.

3. Summary

With the described exercises, students have been trained in using a specialized software for knowledge representation, dealing with problems on the lexical level (reviewing or assigning Basque lexicalizations to concepts), and on semantic level (reviewing or assigning semantic relations such as hyponymy or metonymy), having to deliver supporting evidence for every terminological decision from Basque scientific articles or other sources. They have also documented their project starting point, methodology, workflow, and results on a public page, which can be subject to

¹⁹ See <https://eusterm.wikibase.cloud/wiki/Item:Q9635> for an example entry, with English terms and definitions from different sources.

a collective review process. Last but not least, they have seen themselves empowered as contributors to a large open knowledge graph, where they have added data in their working language, for the benefit of all users.

References

- Cabré, M. T. (2000). La enseñanza de la terminología en España: Problemas y propuestas. *Hermeneus: Revista de la Facultad de Traducción e Interpretación de Soria*, 2, 41–94.
- Costa, R. (2013). Terminology and Specialised Lexicography: Two complementary domains. *Lexicographica*, 29(1), 29–42. <https://doi.org/10.1515/lexi-2013-0004>
- Fontenelle, T., & Rummel, D. (2014). Term banks. In P. Hanks, & G.-M. de Schryver (Eds.), *International Handbook of Modern Lexis and Lexicography* (pp. 1–12). Springer. https://doi.org/10.1007/978-3-642-45369-4_21-1
- Martín-Chozas, P., Ahmadi, S., & Montiel-Ponsoda, E. (2020). Defying Wikidata: Validation of Terminological Relations in the Web of Data. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 5654–5659). European Language Resources Association. <https://aclanthology.org/2020.lrec-1.694>
- Reineke, D., & Romary, L. (2019). Bridging the gap between SKOS and TBX. *Edition – Die Fachzeitschrift Für Terminologie*, 2–2019, 19–27.
- Vrandečić, D. (2012). Wikidata: A new platform for collaborative data collection. *Proceedings of the 21st International Conference on World Wide Web* (pp. 1063–1064). <https://doi.org/10.1145/2187980.2188242>
- Waagmeester, A., Stupp, G., Burgstaller-Muehlbacher, S., Good, B. M., Griffith, M., Griffith, O. L., Hanspers, K., Hermjakob, H., Hudson, T. S., Hybiske, K., Keating, S. M., Manske, M., Mayers, M., Mietchen, D., Mitraka, E., Pico, A. R., Putman, T., Riutta, A., Queralt-Rosinach, N., & Su, A. I. (2020). Wikidata as a knowledge graph for the life sciences. *eLife*, 9, e52614. <https://doi.org/10.7554/eLife.52614>

Acknowledgements

The research leading to the presented results has been supported by DLTB research group (Basque Government, IT1534-2) and Monumenta Linguae Vasconum research project (MINECO, PID2020-118445GB-I00, Govt. of Spain).

Contact information

David Lindemann

UPV/EHU University of the Basque Country
david.lindemann@ehu.eus

Bruno Nahod

CAN WE SUBSTITUTE FIELD EXPERTS WITH CUSTOMIZED LARGE LANGUAGE MODEL IN PROCESSING SPECIALIZED LANGUAGES? A Case Study

Abstract After the Croatian national termbase Struna ceased to receive funding in 2019, we began developing a novel model for compiling terminological collections that will not rely on field experts to provide initial terminological information. A potential solution to our issue of finding a practical and dependable source for obtaining information in the initial stages of processing terminology (i.e., the ‘raw definitions’) across multiple domains could be the publicly available AI language model developed by OpenAI known as GPT-4. GPT is a substantial language model that offers a range of capabilities, including answering queries, generating text, and executing tasks like translation and summarization.

A custom GPT is currently being devised as an aid module, delivering unprocessed information for terminological units that will be processed in Struna. The initial training phase involved manually providing guidelines for best practices in terminology management, which were designed based on the well-established and successful methodology we used to train field experts in the past. The second phase involves feeding TermAI with modified data that was exported from Struna.

In this paper, we will present the results of the comparative analysis of generated terminological units from TermAI and field experts in the domain of forensic sciences.

Keywords artificial intelligence; GPT-4; terminology management: Struna; definition generation

1. Introduction

In the context of terminology management in small languages like Croatian, the importance of field experts in creating high-quality definitions and hierarchical structures has been frequently stressed. In Struna (Struna, 2024), the Croatian national termbase, the role of field experts has been established as a crucial aspect of the workflow (Bratanić & Ostroški Anić, 2015). Their participation has significantly contributed to establishing and upholding standards for terminological collections in Struna. Due to unexpected combinations of circumstances, the original model of processing terminology in Struna is no longer a viable option, and we were forced to start developing a new model. A new model (yet to be finalized) will no longer include field experts as the main generators of information. To maintain the standard of quality, field experts will be used, but their role will be shifted more towards consultations during the final stages of processing.

A potential resolution to our challenge of locating a concise and resilient resource for producing information in the initial phases of processing terminology across different domains could potentially be discovered in GPT-4 (OpenAI GPT-4, 2024.). GPT-4 is a family of models that uses deep neural networks to produce natural text (Rees & Lew, 2024) developed by OpenAI. This extensive language model is capable of tasks like responding to queries, generating text, and executing activities such as translation and summarization. The efficiency of GPT-4 in activities that can be comparable to various elements of terminology processing has been supported by studies showing it generating abstracts (Gao et al., 2023; Ma et al., 2023) as well as creating quality essays on different topics (Bašić et al., 2023; Deniko et al., 2015; Nguyen, 2023; Susnjak, 2022). It has been shown that GPT-4 can, with additional external knowledge, be utilized to successfully comprehend cause-effect relationship of Croatian Chakavian dialect (Perak et al., 2024). Furthermore, studies on using GPT to assist in lexicography (Jakubiček & Rundell, 2023; Lew, 2024) have shown that it can be successfully used in various tasks while showing low usefulness in others. The identification of senses is by far the weakest element of the tool's performance (Rees & Lew, 2024). Furthermore, Jakubiček & Rundell (2023) called attention to the difficulty GPT-4 has with figurative elements of the general language. The consensus seems to be that GPT can be a useful lexicography tool if its limitations are recognized and compensated accordingly (de Schryver, 2023). In the context of terminology management, the utilization of AI, such as GPT-4, can offer a potential solution to the challenges faced when processing terminology across different domains (Shahriar & Hayawi, 2023).

We have hypothesized that the utilization of AI in terminology management can play a pivotal role in overcoming the challenges faced in small languages like Croatian (Nahod et al., 2017). By utilizing AI, terminology management processes can be automated to a certain extent, reducing the reliance on field experts and ensuring consistency and accuracy in definitions and hierarchical structures. Our assumption is based on the observation that terminology processing seems to be exempted from just about all tasks that were recognized as problematic for GPT-4 in lexicography (de Schryver, 2023). The concepts of senses and figurative use are not part of the classical approach to terminology processing (Nahod, 2011). The tasks that GPT-4 exceeds, such as answering direct questions, summarizing, and translating, do appear to be closely matched to tasks the field experts had in the workflow of Struna. Specifically, giving exact and concise definitions for a given concept, recognizing hierarchical relations, and providing translation equivalents (Rees & Lew, 2024).

In this study, we conducted a comparative analysis between the outputs generated by TermAI and those provided by human experts in the domain of forensic sciences. This analysis focused on two key aspects: the form of definitions and the semantic accuracy of the generated content. Our investigation led to the formulation of the following hypotheses:

Hypothesis 1

H1: TermAI will perform at a comparable level of accuracy to domain experts in the form of definition test.

H1a: There will be no significant difference in the accuracy of the form of definition test results between queries given in English and those given in Croatian.

H1b: There will be no significant difference in the accuracy of the form of definition test results across different subfields within the domain of forensic sciences.

Hypothesis 2

H2: TermAI will perform at a comparable level of accuracy to domain experts in the semantic match test and the equivalent match test.

H2a: There will be no significant difference in the accuracy of the semantic match test and the equivalent match test results between queries given in English and those given in Croatian.

H2b: There will be no significant difference in the accuracy of the semantic match test and the equivalent match test results across different subfields within the domain of forensic sciences.

These hypotheses reflect our expectation that TermAI, through its customization and training, will match the performance of human experts in most cases. However, the results of this study will also highlight areas where further refinement of the TermAI may be necessary, particularly in specialized or interdisciplinary subfields. Our ultimate goal is to implement GPT-4 into our workflow, where it would be used as the main generator of 'raw definitions', conceptual relations, and translation equivalents in the early stages of the workflow, mimicking the tasks performed by the field experts in the late model. The information generated by the GPT-4 would be processed by Croatian language experts and trained terminologists, followed by conformational editing performed by field experts.

2. Materials and Methods

Following the intensive stage of testing the capabilities and limitations of the general GPT-4, we have initiated the development of custom GPT (GPTs OpenAI, 2024) that allow the customization of the GPT for specific purposes by following the list of personalized prompts and instruction sets defined by the user.

We are currently developing a single purpose custom GPT that we have named TermAI. OpenAI is offering GPTs service (GPTs-OpenAI, 2024) that allows users to customize GPT-4. Where one can, relatively simply, customize a GPT to their own specifications for specific purposes. The main purpose of TermAI is to give a definition and term equivalents for the prompt given in the proscribed form that we are calling a query.

The development and customisation of the TermAI is planned to go through three main stages: defining the rules, manual training, and finally batch training. Naturally, all the stages of customization include multiple revisions by both trained terminologists and our project coordinator. The first stage of customization, which included the

logical and theoretical parts of the training, was based on the workshops we used to train field experts over the last 15 years. During this phase, we have defined the Cardinal Rules—a set of instructions and directives that can't be sidestepped or broken. Currently, there are 9 rules that TermAI must follow when interacting with a user. Most of them concern the proper form of the terminological definition, the structure of the answers, and the acceptable form of the queries. For example, Rule 1: “definition must be one sentence in the form of *genus proximum et differentia specifica*”; Rule 3: “any additional information about the term or the concept must be in the “napomena/note” part of the answer”; and Rule 6: “acceptable forms of the query are; term, term/domain, term/language and term/domain/language.”

Stage two of the training covers the practical part of producing ‘raw definitions’ and is performed by training TermAI with a term-definition form of information sourced from Struna. We decided that the first milestone would be 1000 terminological units from the domain of forensic sciences (Table 1). The domain of forensic sciences was chosen for two reasons: it was the last project in the old model, so most of the team members were familiar with terminological collection and were able to pick the “best” terminological units for training TermAI. Secondly, forensics science was by far the most interdisciplinary collection we have processed in Struna so the terms and definitions processed have a wide semantic field, covering over 20 domains of knowledge.

Table 1: Examples of terminological data from Struna used in the second stage of TermAI training

Croatian term	English term	Definition
poligrafsko ispitivanje	polygraph testing	ispitivanje osobe i praćenje njezinih fizioloških reakcija s pomoću poligrafa radi utvrđivanja istinitosti iskaza 'the testing of a person while monitoring their physiological reactions using a polygraph to determine the truthfulness of their statement'
očevidac	eyewitness	osoba koja je svojim osjetilima neposredno opažala počinjenje kaznenoga djela 'person who directly perceived the criminal offense with their senses'
petlja	loop	crtež papilarnih linija s jednom deltom u kojemu linije teku u smjeru središta crteža gdje se uvijaju i vraćaju natrag u smjeru odakle su došle 'fingerprint pattern with one delta, in which the ridges flow towards the center of the pattern, where they loop and return in the direction from which they came'

After the first milestone of 1000 terminological units was reached, we initiated a series of tests to evaluate the current state of TermAI and identify problems and areas we need to address in future training. Following the promising results of the first series of tests, we designed a study to test TermAI performance. TermAI was given the task of generating definitions for the 20 new concepts from the domain of forensic sciences. We have selected concepts that were not previously processed in Struna and have not been used in TermAi's training so far.

The study was designed as a two-group comparative, where the same terminological units were generated by forensic sciences experts.

As a source for our concepts a small corpus was compiled using 150 thesis papers provided by the University Department of Forensic Sciences, University of Split. The extracted term candidates were filtered against the terms edited in Struna and the final collection of 20 terms were selected to be used in this analysis. Considering how forensic sciences is extremely interdisciplinary we have tried to select the concepts from the broader domain of forensic sciences that closely represented the semantical broadness of the terminological units processed in *Forensic and Criminal Investigation Project* FuNK (Bašić, 2024). Selected terms varied on multiple levels: single-word ($n = 7$), multi-word ($n = 13$); by subfields law and legislation ($n = 7$), security and defence sciences ($n = 8$), information science ($n = 2$), and sociology, ethics, psychiatry one ($n = 1$) each.

The TermAI was given queries in the form of ‘term/domain’ which is the form defined by one of the rules that state that for this form of a query, it should give the response representing the specific semantic field corresponding to the domain in the query, both in English and Croatian. Queries were divided into two sets: set A had all the terms in the Croatian language and set B had all the terms in English (Table 2).

Table 2: Queries by set

Set A	<i>neprihvatljiv dokaz</i> ‘inadmissible evidence’, <i>teško ubojstvo</i> ‘aggravated murder’, <i>digitalni dvojniki</i> ‘digital twin’, <i>krunski svjedok</i> ‘crown witness’, <i>posebna dokazna radnja</i> ‘special investigative measure’, <i>ugrožena osoba</i> ‘endangered person’, <i>kvalificirano kazneno djelo</i> ‘serious crime’, <i>kriptografija</i> ‘cryptography’, <i>nedominanta ruka</i> ‘non-dominant hand’, <i>čestitost</i> ‘integrity’
Set B	treachery, grooming, para-suicidality, socioeconomic status, hybrid warfare, hybrid threat, national security, trinitrotoluene, espionage, counterintelligence

The main goal was to evaluate TermAI’s ability to generate definitions in the domain in which was trained. That would, hopefully, give us a better understanding of its current state of development and allow us to identify its weaknesses. Our secondary goal was to determine the quality of term translation, both from Croatian to English and from English to Croatian.

For the purpose of this study, we formed a group of forensic sciences experts ($n=3$), all of whom were previously involved in the FuNK project and had undertaken a full terminological and linguistic training 5 years ago. They were asked to provide definitions and English equivalents for the same 20 concepts. For the ease of the data analysis, we have named them Group A, while TermAI was named Group B.

3. Results

The data generated by Group B (TermAI) was evaluated and scored against the data generated by Group A (Table 3). Generated terminological units were parsed

into three categories: **equivalent match** – where we were looking if the generated Croatian and English equivalents correspond to their pairs from the Group A, **semantic match** – where we were looking how closely the definition corresponds to the definitions from the Group A, and **form score** – where each definition was scored based on the terminological principles enforced in Struna. And finally, following the same rules, we have also scored the definitions generated by the Group A. The scoring for the **form** was 1–5, where 5 was a perfect definition, 4 needing minimal terminological and/or language intervention, 3 – a considerable intervention is needed, 2 – definition needs a full rewrite with consultations, and 1 – a new definition is needed.

Table 3: Values by category used in evaluation of the data

Equivalent match	Semantic match	Form score
Perfect match	Match	1–5
Synonym	Different aspect	
Fail	Fail	

The first variable analysed to test Hypothesis 1 was the terminological form of definitions produced by both Group A (domain experts) and Group B (TermAI). The descriptive statistics for the form scores across both groups are summarized below:

Statistic	Group A form score	Group B form score
Count	20	20
Mean	4.42	4.84
Standard Deviation	0.77	0.37
Min	3.00	4.00
25th Percentile	4.00	5.00
50th Percentile (Median)	5.00	5.00
75th Percentile	5.00	5.00
Max	5.00	5.00

To evaluate the difference in form scores between Group A and Group B, an independent t-test was conducted. The results were as follows:

T-statistic: -2.147

P-value: 0.039

Given that the p-value is less than the significance level of 0.05, we can conclude that there is a statistically significant difference between the form scores of Group A and Group B. Specifically, TermAI (Group B) demonstrated a higher mean form score compared to the human experts (Group A), indicating that TermAI was able to produce terminological definitions that adhered more closely to the standardized form criteria.

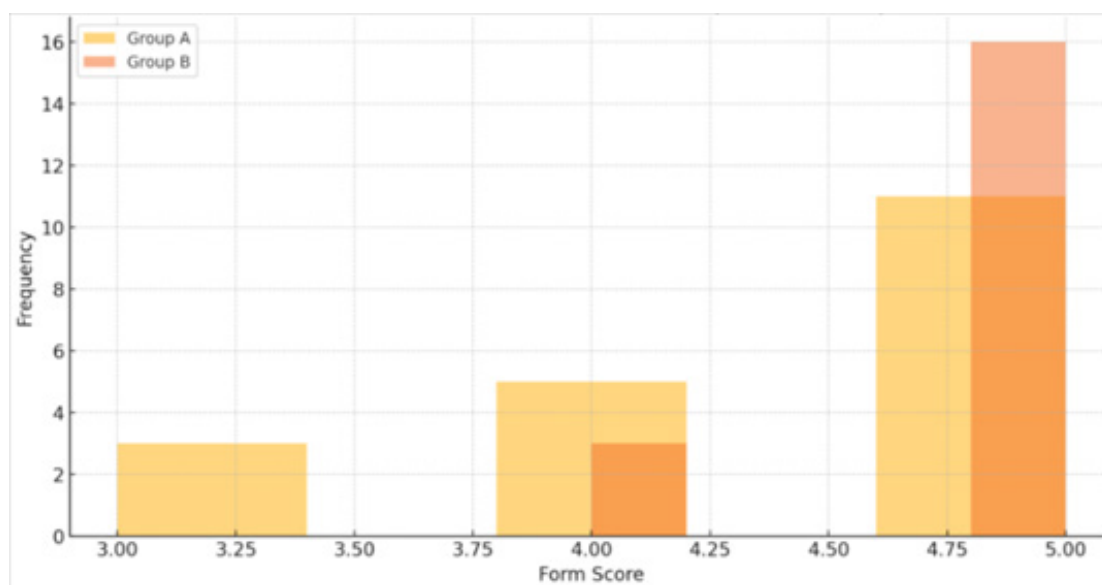


Fig. 1: Distributions for the **form score** for Groups A & B

The overall distribution analysis of the form scores for TermAI indicates a high level of terminological accuracy, with all scores falling within the 4 to 5 range on a 1–5 scale. This suggests that TermAI consistently produced well-formed definitions. A slight difference in performance was observed between queries given in Croatian and those in English, with the Croatian queries achieving a 90% ratio of scores at 5, compared to an 80% ratio for English queries. This observation is supported by a weak negative correlation (-0.140) between the query language and the form score, suggesting only a minimal impact of language on TermAI's performance.

To test Hypothesis 2 and assess whether TermAI performs at a comparable level of accuracy across different subfields, we analyzed the correlation between the Equivalent Match Score and the Semantic Match Score with the various subfields included in this study. The correlations are summarized below:

Subfield	Correlation with Equivalent Match Score
IT	0.249
SDS	0.440
law	-0.815
rest	0.316
Subfield	Correlation with Semantic Match Score
IT	-0.094
SDS	0.131
law	-0.095
rest	0.031

The data reveal a notable negative correlation between the law subfield and both the Equivalent Match Score (-0.815) and the Semantic Match Score (-0.095). This suggests that TermAI's performance is less accurate in the law subfield compared to others, indicating that higher scores are less likely to be associated with legal terminology.

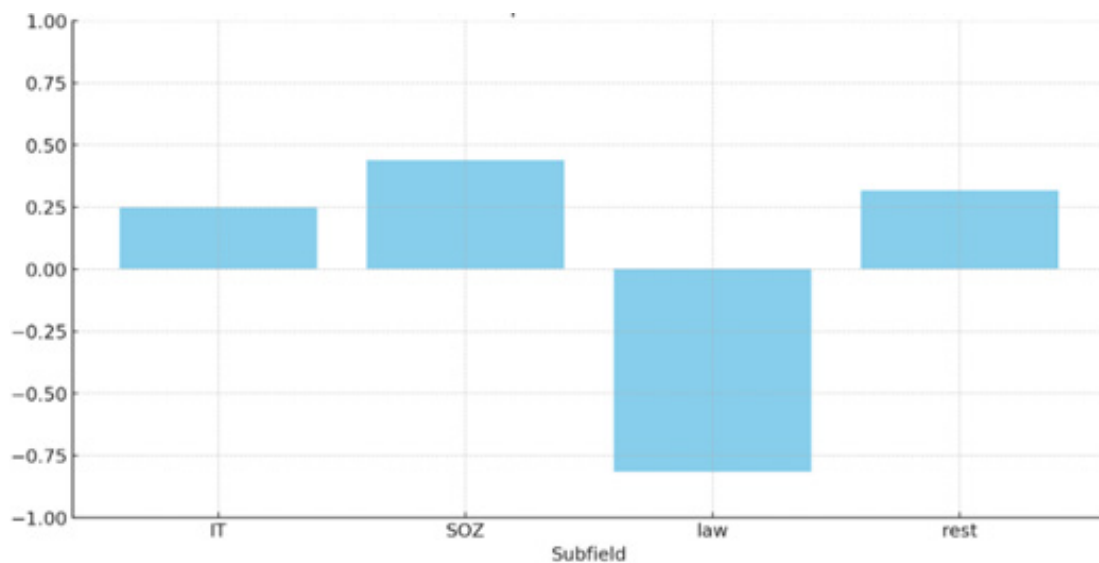


Fig. 2: Correlation between Equivalent match and Subfield



Fig. 3: Correlation between Semantic match and Subfield

A quantitative analysis of the “Equivalent Match” for each “Query Language” was conducted to further explore the performance of TermAI, particularly in relation to Hypothesis 2. The analysis revealed the following distributions:

- Croatian queries: Fail: 20%, Original 0%, Perfect match 50%, Synonym 30%
- English queries: Fail 10%, Original 20%, Perfect Match 70%, Synonym 0%.

The results indicate that for Croatian queries, TermAI produced a perfect match for 50% of the cases, while for English queries, the perfect match rate was higher at 70%. Conversely, the failure rate was higher for Croatian queries (20%) compared to English queries (10%).

A weak negative correlation (-0.169) between the “Equivalent Match” and the language of the query suggests a slight tendency for the accuracy of equivalent matches to vary with the language. Specifically, as the query language changes from Croatian to English, the likelihood of achieving a perfect or original equivalent match slightly decreases in its numerical value.

4. Discussion

As far as we are aware, this is the first study that tested ChatGPT as a terminology unit generating tool in comparison to human field experts. Our study showed that TermAI preformed beyond expectations for its current state of development. Compared to (subjectively) best terminologically trained field experts in the domain of forensic sciences, TermAI performed on the level that would be acceptable for the terminology processing in Struna. With the mean score of 4.85 for **form score** which represents the level of acceptancy in the definition form it slightly overperformed trained domain experts (mean 4.42).

Notably, the form of the definitions is not the most important aspect of terminology processing, but given our experience reaching this level of performance has generally been a struggle while training domain experts. The accuracy of definitions which were presented with the **semantic match** variable does show observable problems in matching domain experts. Of the 20 definitions generated by TermAI, 13 (65%) definitions that were semantically identical to the corresponding definition given by the domain experts, as far as we were able to judge. The remaining 7 (35%) were not wrong *per se*, but were generated with a noticeable semantic shift. The majority of them (42%) were from the subfield of law and legislation, the results when combined with a negative correlation (-0.815) for **equivalent match** with the same subfield does show that TermAI has performance problems in that subfield when dealing in interdisciplinary domain of forensic sciences. Considering the nature of the law and legislation terminology, which can be described as highly constricted and dependent on specific aspects such as country, tradition, and case-by-case application (Fajfar et al., 2019), this should not come as a surprise.

The analysis of the term equivalents generation TermAI has also performed well with 60% of generated equivalents being an exact match to the ones provided by domain experts. In 3 cases (1.5%) TermAI generated acceptable synonyms, results which we cannot but count as accurate. The rest of the generated equivalents consist of 3 errors, where TermAI generated wrong terms of those 2 were in English and 1 in Croatian, all of them in the subfield of law and legislation: *key witness* for *crown witness*, *vulnerable person* for *endangered person* and *izdaja* for *podmuklost*. Notably the Croatian term *podmuklost* which was given by the domain experts for the English term *treachery* is a strange choice, and (subjectively) we would consider *izdaja* to a better translation. The remaining two equivalents TermAI left in their original English form when generating Croatian terms, which could be considered a better solution than generating a wrong term. Notably, both Croatian terms provided by forensics experts (*mamljenje* for *grooming* and *spufiranje* for *spoofing*) are recent

reterminologization and newly coined ones respectively. Therefore, it should not come as a surprise that TermAI wasn't able to provide a matching term equivalent. During this study TermAI hasn't presented a single case of "hallucinations", the fact that we could correlate to the strict and comprehensive theoretical training we provided at the very early stages of development.

5. Conclusion

The outcomes of our study surpassed our expectations. Despite anticipating significant challenges in the early stages of training, our custom GPT TermAI demonstrated remarkable performance. Achieving an overall accuracy exceeding 60% in generating both definitions and term equivalents was more than we initially anticipated at this stage. Notably, if we consider that most of the excluded generated definition-equivalent answers are not fundamentally incorrect but rather not immediately usable, the percentage of potentially useful 'raw definitions' and term equivalents can subjectively be elevated to 90%.

The objectivity of our analysis warrants attention. Evaluating any written information is inherently subjective, especially when dealing with specialized knowledge meanings and definitions. Acknowledging this, we consciously designed the study with certain controls. Definitions generated by TermAI, intended as 'raw definitions' for further refinement by trained terminologists, were compared to those provided by domain experts. These experts, trained by our team, also had experience working on the FuNK project in Struna. Typically, we receive 'raw definitions' from domain experts in fields outside our expertise, such as physics, engineering, and forensic sciences. For this study, we intentionally selected the domain of forensic sciences due to our team members' involvement in the FuNK project, enabling us to promptly verify the accuracy of generated answers. We can consider this a deviation from standard practice, which enabled us to be stricter while evaluating definitions. Considering this bias, the results are even more encouraging.

The subsequent phase involves training TermAI on 20,000 terminological units provided in domain-based batches. This will follow additional studies aimed at identifying limitations and challenges with our custom GPT. We hope that these insights will refine TermAI's capabilities and introduce new features. Our results suggest that additional effort is required to enhance TermAI's handling of interdisciplinary concepts, particularly in integrating specific semantic aspects of definitions within assigned domains. Moreover, it is clear that a synonyms section should be incorporated into TermAI's answers to avoid potential issues for future users. The analysis also indicates a need to improve TermAI's term translation abilities, especially for English-language queries.

At this developmental stage, the most significant progress is observed in the quality of standardized generated definitions. While ChatGPT is generally unpredictable in both form and content, our efforts in training and implementing Cardinal Rules have resulted in a bot that is relatively stable and predictable.

AI models like GPT-4 are paving new paths in language research. Despite the inherent risks of over-reliance, these models can be valuable tools when their capabilities and limitations are understood. At this stage, despite TermAI's promising performance, we are hesitant to grant it full autonomy in user interactions. We hope that continued research and training will enable us to develop TermAI to a level where it can effectively complement the content of Struna.

And finally, to answer the question from the title – Can we substitute field experts with AI in early stages of specialized languages processing? The results of this study suggests that at this stage we cannot fully depend on AI to substitute real human experts. On the other hand, the results are promising with strong implication that with further research and training AI models could be turned into a useful tool in terminology processing for Croatian language.

References

- Bašić, Ž. (n.d.). HRZZ - STRUNA - FuNK. Retrieved June 1, 2024, from <https://forenzika.unist.hr/forenzicno-kriminalisticko-nazivlje-struna-funk/>
- Bašić, Ž., Banovac, A., Kružić, I., & Jerković, I. (2023). ChatGPT-3.5 as writing assistance in students' essays. *Humanities and Social Sciences Communications*, 10(1), 1–6. <https://doi.org/10.1057/s41599-023-02269-7>
- Bratanić, M., & Ostroški Anić, A. (2015). Konceptija i ustrojstvo terminološke baze Struna. In M. Bratanić, I. Brač, & B. Pritchard (Eds.), *Od Šuleka do Schengena: Terminološki, terminografski i prijevodni aspekti jezika struke* (pp. 57–74). Institut za hrvatski jezik i jezikoslovlje.
- Deniko, R. V., Shchitova, O. G., Shchitova, D. A., & Lan, N. T. (2015). Learning Terminology in the Age of Higher Education Internationalization: Problems and Solutions. *Procedia - Social and Behavioral Sciences*, 215(June), 107–111. <https://doi.org/10.1016/j.sbspro.2015.11.582>
- de Schryver, G.-M. (2023). Generative AI and Lexicography: The Current State of the Art Using ChatGPT. *International Journal of Lexicography*, 36(4), 355–387. <https://doi.org/10.1093/ijl/ecad021>
- Fajfar, T., Tomazin, M. J., & Karer, M. Z. (2019). Slovenian legal terminology and its presentation in the dictionary of legal terminology. *Jezikoslovni zapiski*, 25(1), 53–66. <https://doi.org/10.3986/jz.v25i1.7565>
- Gao, C. A., Howard, F. M., Markov, N. S., Dyer, E. C., Ramesh, S., Luo, Y., & Pearson, A. T. (2023). Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. *Npj Digital Medicine*, 6(1), 1–5. <https://doi.org/10.1038/s41746-023-00819-6>
- GPTs-OpenAI. (2024). Introducing GPTs. <https://openai.com/blog/introducing-gpts>
- Jakubiček, M., & Rundell, M. (2023). The end of lexicography? Can ChatGPT outperform current tools for post-editing lexicography? In I. Kosem, M. Cukr, M. Jakubiček, J. Kallas, S. Krek, & C. Tiberius (Eds.), *Proceedings of Electronic Lexicography in the 21st Century Conference* (pp. 518–533). Lexical Computing CZ, s.r.o.

Lew, R. (2024). Dictionaries and lexicography in the AI era. *Humanities and Social Sciences Communications*, 11(1), 1-8. <https://doi.org/10.1057/s41599-024-02889-7>

Ma, Y., Liu, J., Yi, F., Cheng, Q., Huang, Y., Lu, W., & Liu, X. (2023). AI vs. Human -- Differentiation Analysis of Scientific Content Generation. January. <https://doi.org/10.48550/arXiv.2301.10416>

Nahod, B. (2011). Od polisemna leksema do homonimna naziva. In L. Pon, V. Karabalić, & S. Cimer (Eds.), *Aktualna istraživanja u primijenjenoj lingvistici. Zbornik radova s 25. međunarodnog skupa HDPL-a održanog 12.-14. svibnja 2011. u Osijeku* (pp. 15-26). Hrvatsko društvo za primijenjenu lingvistiku.

Nahod, B., Nahod, P. V., & Bjeloš, M. (2017). Three aspects of processing ophthalmological terminology in a "small language": a case of Croatian term bank Struna. *Jazykovedny casopis*, 68(2), 287-295. <https://doi.org/10.1515/jazcas-2017-0038>

Nguyen, M. H. (2023). Academic writing and AI: Day-1 experiment. Center for Open Science 2023, January. <https://osf.io/kr29c/download>

OpenAI. (2024). Introducing GPTs. <https://openai.com/blog/introducing-gpts>

OpenAI GPT-4. (n.d.). <https://openai.com/product/gpt-4>

Perak, B., Beliga, S., & Meštrović, A. (2024). Incorporating Dialect Understanding Into LLM Using RAG and Prompt Engineering Techniques for Causal Commonsense Reasoning. In Y. Scherrer, T. Jauhainen, N. Ljubešić, M. Zampieri, P. Nakov, & J. Tiedemann (Eds.), *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects* (pp. 220-229). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.vardial-1.19>

Rees, G. P., & Lew, R. (2024). The Effectiveness of OpenAI GPT-Generated Definitions Versus Definitions from an English Learners' Dictionary in a Lexically Orientated Reading Task. *International Journal of Lexicography*, 37(1), 50-74. <https://doi.org/10.1093/ijl/ecad030>

Shahriar, S., & Hayawi, K. (2023). Let's Have a Chat! A Conversation with ChatGPT: Technology, Applications, and Limitations. *Artificial Intelligence and Applications*, 2(1), 11-20. <https://doi.org/10.47852/bonviewaia3202939>

Struna. (2024). Struna. <http://struna.ihjj.hr/>

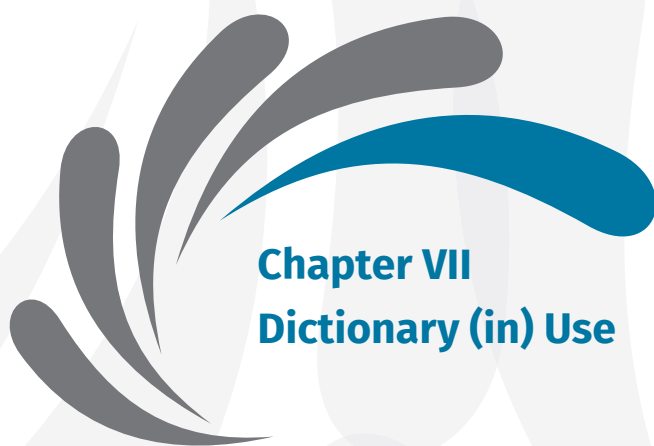
Susnjak, T. (2022). ChatGPT: The End of Online Exam Integrity? 1-21. <http://arxiv.org/abs/2212.09292>

Acknowledgements

This paper was created within the framework of the project TermAI: Development and Modernization of Croatian Terminology, which is funded by the European Union – NextGenerationEU. The views and opinions expressed are solely those of the authors and do not necessarily reflect the official positions of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them.

Contact information

Bruno Nahod
Institute for the Croatian Language
bnahod@ihjj.hr



Chapter VII
Dictionary (in) Use

Margit Langemets, Lydia Risberg, and Kristel Algvere

TO DREAM OR NOT TO DREAM ABOUT 'CORRECT' MEANINGS? Insights into the User Experience Survey

Abstract In 2023, the Institute of the Estonian Language, in collaboration with the Center for Applied Anthropology of Estonia, conducted a user experience survey aimed at understanding the habits, needs, and attitudes of users of the language portal *Sõnaveeb* ('Word Web') and preparing for the publication of the *Dictionary of Standard Estonian* (DSE) in 2025. This paper addresses prescriptive and descriptive issues in Estonian lexicography, including controversial meanings. It provides an overview of the user experience survey, detailing the methodology of the online survey and the subsequent qualitative analysis of responses. The findings from the survey are presented and discussed, revealing diverse attitudes toward language among users. While some dictionary users only seek information about language, others aim to enrich their language use, and a third group seeks correctness and guidelines from language planning, including 'correct' meanings of words. The contemporary linguistic approach is descriptive rather than normative and this has been adopted in the *EKI Combined Dictionary* (available via the language portal *Sõnaveeb*). However, the legal norm of Standard Estonian is still determined (among other sources) by the latest *Dictionary of Standard Estonian* (DSE). The existence of these two separate sources has been causing confusion among dictionary users.

Keywords: user experience survey; Estonian; prescriptive and descriptive lexicography; attitudes

1. Introduction

Since 2019, the Institute of the Estonian Language (EKI) has been developing the language portal *Sõnaveeb* ('Word Web', Tavast et al., 2018; Koppel et al., 2019). Increasing amounts of data from various general dictionaries have been integrated into the *EKI Combined Dictionary* (CombiDic), displaying information from previous EKI dictionaries: explanatory (2019), collocation (2019), Standard Estonian (DSE 2018), bilingual, etc. Besides the general language lexical data, over 130 multilingual terminological dictionaries are available via the same portal, altogether comprising nearly 400,000 Estonian headwords. The number of users has risen from 10,000 in January 2020 to its highest in March 2024, reaching 112,000 users (according to Google Analytics). By device category, the mobile is preferred (54%) over the desktop (45%).

The Institute has been commissioned by legislative acts (Language Act, 2011; Regulation 2011) to ensure the availability of information about Standard Estonian, fixing the standards and recommendations of the official and public use of language. The norm of Standard Estonian is determined (among other sources) by the latest *Dictionary of Standard Estonian* (DSE), published twice since 2011 (DSE 2013, DSE 2018). The next

DSE is to be published in 2025. The DSE plays a significant role in Estonian society, especially for copy editors, teachers, and other language professionals. The DSE has enjoyed not only societal prestige but also legal support (in the 1920s-1930s see Tauli, 1943; since the 1990s see the Language Act, 2011).

Since 2020, Estonia has faced a quite heated discussion on whether the approach to language and lexicography should be descriptive or prescriptive (see e.g., Vainik & Paulsen, 2023; Koreinik, 2023; Lindström, Risberg, & Plado, 2023, pp. 32–37). In March 2023, the *Action Plan* was launched by the Ministry of Education for preparing the DSE 2025, then to be published in the language portal *Sõnaveeb*, also serving the need for clarification of relevant language resources and services. As the “dictionary war” advanced rather than retreated, the *Action Plan* was updated in May 2024. The result is that the DSE will continue to be published on paper as previously, together with its static electronic copy on the separate EKI web page. The DSE 2025 will serve as a basis for the literary norm of Standard Estonian from January 1st, 2026.

In 2023, the user experience survey was conducted by EKI, in collaboration with the Center for Applied Anthropology of Estonia (CAAE) to better understand the habits, needs, and attitudes of the users of the language portal *Sõnaveeb*, as well as to prepare the publication of the DSE 2025 (Langemets et al., 2024b). The different attitudes that emerged inductively from the comments to the open-ended questions in the study allow mapping patterns of attitudes towards lexicographic support and give an insight into the attitudes of survey participants. As language codifiers such as the DSE and *Sõnaveeb* are one of the social forces that determine what is standard in a language (Ammon, 2015, p. 57), it is essential to study the attitudes towards them. Attitudes are at the heart of the human experience and play a pervasive role in our daily interactions (e.g., Fazio & Olson, 2003), influencing how we present ourselves and interact with others. Language users’ attitudes and opinions are shaped, among other factors, by personal experience, social context, and media (Garrett, 2010, p. 22).

In this paper, we first touch upon normative and descriptive issues in Estonian lexicography (1.1), including controversial meanings (2.2). In Section 2, we give an overview of the user experience survey, describing the methodology of the online survey (2.1) and the analysis of the answers (2.2). In Section 3 we present the results of the survey, followed by a discussion in Section 4. We will end with a conclusion in Section 5.

1.1 Prescriptive and Descriptive Lexicography

The Estonian Literary Standard, as stated in the law, and to be presented in the DSE is meant to be ‘the system of spelling, grammar and lexical standards and recommendations’ regulating the official use of language (Language Act, 2011). In other texts directed to the public, including the use of Estonian in the media, the best practice of the use of language is to be followed. While the terms *spelling* (orthography) and *grammar* (inflection) seem to be commonly understood as well as subjected to normative purposes, *lexical* is not. Typically, it means just ‘relating to the words’, i.e., it does not imply meanings unavoidably, but, instead word formation, constructions,

or maybe word choice in different contexts. However, during public discussion, the aspect of the normativeness of word meanings has been raised. It should be mentioned also, that in language planning, the theoretical basis for providing recommendations about word meanings has been lacking (Risberg, 2024, p. 98).

Substantial monolingual dictionaries typically list the specific number of meanings in which a word is used, in some cases citing up to 20 senses or more. This has not been the case in the DSE, as historically, since the first DSE in 1918, it has followed the example of the *Duden Rechtschreibwörterbuch* (9th edition in 1915), enacting foremost the rules for orthography and inflection and only some semantic comments (for neologisms and loanwords). In the course of time, more and more hints on meanings have been added to the DSE. Though it never formally addressed the presentation of *all* the meanings, in 1961 the compilation of the explanatory dictionary of Estonian (EKSS) was initiated (the first fascicle was published in 1988). Regarding the DSE, a shift took place in the 1980s, transmitting the DSE 1999 into a *language planning* dictionary offering advice – not rigid rules – on ‘correct’ (or ‘good’) and ‘incorrect’ (‘bad’) use of language. (However, the recommendations were phrased in a way that dictionary users might not have realized they were not mandatory rules.)

Since the establishment of legal regulations in 2011, the DSE has developed into a ‘legal norm of a special kind’ that has little to do with empirical findings in linguistics. Although meanings are not mentioned in the regulations and recent research has proved once again that word meanings cannot be standardized without these prescriptions becoming irrelevant (Risberg, 2024), it is not at all clear among the Estonian community whether (all or some) meanings are part of the legal norm or not. The *Action Plan* (2024) envisages for the DSE 2025 a ‘sameness’ with DSE 2011 and 2013 – the dictionaries lacking many meanings – which leaves us the unsolved problem of whether the meanings are subject or not to the legal norm. The discussion in the community has raised even further questions: does the dictionary violate the *Language Act*, describing actual language use?

Since the 1870s, which is considered the beginning of language planning in Estonia (T. Ereht, 2002), the Estonian language community has been strongly influenced by the Standard Language Ideology (e.g., Milroy, 2001), a belief in the existence of a superior language variant as well as the pursuit of purism, the rejection of undesirable language elements and foreign influences (see Lindström, Risberg, & Plado, 2023). This belief has been strengthened by the existence of two separate dictionaries: one for looking up the norms (normative DSE), the other for searching the actual language data (descriptive EKSS 2009, CombiDic). The history of the explanatory dictionary in Estonia is 70 years shorter than that of the DSE, this might have influenced the understanding of how language works and beliefs about language in the society.

1.2 Controversial Meanings

In line with usage-based linguistics (Diessel, 2017; Díaz-Campos & Balasch, 2023), EKI has moved towards the descriptive approach in language planning, and regarding meanings,

we have applied the principle that meanings of general language are not standardized (Risberg, 2024). The dictionary can provide additional descriptive comments on specific usage cases, such as discriminatory language, but it cannot provide recommendations for avoiding certain common meanings in any communicative situation.

More than 20 years ago Lars Trap-Jensen (2002) expanded on some borderline cases in Denmark. The Danish Dictionary was commissioned to prepare a dictionary that should be “descriptive in the overall aim. It should, however, be normative with respect to information on orthography, and guiding with respect to information on inflection, word formation, collocations, and constructions” (quote from the official mandate). Regarding controversial meanings, they implemented a label ‘this usage is regarded by some/many as incorrect’ (Trap-Jensen, 2002).

In Estonia, we have decided to include the meanings in CombiDic if they meet the standard criteria for description in descriptive lexicography and, if necessary, provide a special note on usage to consider, particularly in official communication. We have used the wording ‘instead of usage X one might consider using Y or Z in official language’. In general, we would prefer to offer substantiated information for the user to be able to choose between different options, instead of generally stigmatizing some words or word forms as ‘incorrect’.

Since 2019 our goal has been to build a “single-window system” to offer all lexical information from one source for the user. We have integrated the DSE 2018 into the CombiDic (via the language portal *Sõnaveeb*), combining descriptive and prescriptive information, hoping to benefit both, the user seeking information on (actual) language as well as the user seeking guidance on ‘good’ language use. The latest developments, connected with juridical matters, have radically changed the scene: the DSE 2025 is to be published separately (both in print and as a corresponding web copy) and preserved unchanging for the next 7-10 years. As for the content, it should follow the type of DSE 2013 and 2018, so there will be only a handful of meanings, including the notes for the controversial issues. *All* the meanings (together with the notes) will be available in CombiDic, to be updated constantly.

2. User Experience Survey

Different levels of research on dictionary use exist across Europe, ranging from extensive studies in some countries to little or no research in others (Kosem et al., 2018, p. 91; e.g., see Dziemianko, 2014; Lew, 2015a, 2015b; Müller-Spitzer, 2014; Wolfer et al., 2018). The use of dictionaries can be studied in various ways, such as conducting an online survey (Kosem et al., 2019), and using online tracking technologies (Wolfer, Michaelis, & Müller-Spitzer, 2021), among others.

In 2023, the user experience survey was conducted by the Institute (EKI) in collaboration with the Center for Applied Anthropology of Estonia (CAAE) to better understand the habits, needs, and attitudes of the users of the language portal *Sõnaveeb*, as well as to prepare the publication of the DSE 2025. The study consisted

of an online survey prepared by EKI and 16 in-depth interviews (60 minutes each) conducted by CAAE, aiming to get an overview of how the different user groups evaluate their user experience. Our study was focused on the general public – while language professionals were involved, many others also participated (see Section 2.1). Our aim was to involve as many users of the portal as possible to understand how and for what the portal is used and why it is not used, while also allowing participants flexibility in their responses. We used the “what people say”-approach to get to know participants’ stated beliefs they are aware of and willing to report. In this paper,¹ we will analyze the open-ended feedback questions (comments) in the online survey, elaborated and analyzed by the authors. After describing the design and dataset, the analytical approach and coding frame are presented.

The questions in the online survey (see Section 2.1) were not specifically about attitudes, however, they contrasted strongly with the comments. We will also be presenting and analyzing this language dimension in our results as it serves to gain insight into the attitudes and opinions of language users and is an important part of sociolinguistics (Dragojevic et al., 2021; Ebner, 2017, pp. 61–90; Garrett, 2010; Giles & Marlow, 2011, p. 183). The qualitative analysis (Section 2.2) gives insight into the attitudes and linguistic stereotypes of the survey participants.

Linguistic stereotypes have a lot to do with attitudes and these are organized along two major evaluative dimensions: *social status* (e.g., competent, intelligent) and *solidarity* (e.g., warm, friendly, attractive) (Edwards, 1999, p. 102; Dragojevic et al., 2018, p. 30). Attitudes are also shaped by ideologies. Often language users do not realize the influence of Standard Language Ideology but tend to see linguistic standards as a matter of common sense (Garrett, 2010, p. 7). In the Estonian school system, Standard Estonian is often taught as the only correct form of Estonian and is widely considered a central component of the identity of Estonians (Lindström, Risberg, & Plado, 2023, p. 9). The high prestige of standard language in society, including the obligation to use only ‘correct’ language at school, may seriously impact the linguistic security of language users (Vaicekauskienė, 2012, p. 77). Furthermore, when the state institutionalizes overt ideology, it inevitably influences the linguistic identity of speakers. This is particularly true for professional language users who are more exposed to the high prestige of the official language (Vaicekauskienė, 2012, p. 98).

2.1 Online Survey

The online survey (in LimeSurvey) was reachable for about 1.5 months (18.05-30.06.2023). The core survey comprised 21 questions, including 7 about personal data, taking about 10-15 minutes to answer. It was published on EKI’s website, called for in the Estonian Public Broadcasting (“Keeleminutid”, 13. 06. 2023), shared on Facebook, and (as a pop-up) in the language portal *Sõnaveeb*. A total of 3,005 respondents participated in the survey, while 1,382 (45.9%) completed it (Langemets et al., 2024b).²

¹ This paper is partly based on two conference presentations held in Estonia in April 2024 focusing on attitudes.

² Although it was clearly stated in the welcoming text that respondents must be at least 18 years old, the questionnaire had 52 minors to answer. Their responses have been deleted prior to analysis.

The questions spread across seven modules: (1) using the portal *Sõnaveeb* (yes/no, how often, devices, giving feedback); (2) evaluating types of language data; (3) layout of the portal; (4) the medium of the DSE (paper, online); (5) the overall assessment of the portal; (6) personal data (age, gender, formal education, mother tongue(s)); (7) activities when a dictionary consultation is needed.³ Due to the nature of the study, we had to rely on the self-reports of the participants.

The language portal was used (Q1, N=2,151) by 1752 (81.4%), not used by 135 respondents (6.3%), and 264 respondents (12.3%) noted that they heard about it from this questionnaire. The overall satisfaction with the portal (Q3, N=1,333) was high: the majority of respondents were satisfied (1,116 respondents, 83.7%); 153 respondents stayed neutral (11.5%); few were unsatisfied (64 respondents, 4.8%).

Different age groups (Q16, N=1,291) were covered quite evenly: each 5-year range from 18-19 to 50-54 around 10%, then lowering from 6.3% (55-59) to 0.2% (85+). As for gender groups (Q17, N=1,312), 71.2% were women (934 respondents), 26.7% were men (350), and 2.1% marked “Other” or were not willing to tell (28). The data on education (Q18, N=1,265) showed that the majority of respondents were MAs or MA students (44.6%, 565 respondents), followed by BAs or BA students (24.6%, 311), secondary or vocational education (19.4%, 245), PhDs or PhD students (7.6%, 96), and basic education (3.8%, 48). The mother tongue(s) of the respondents (Q19, N=1,413) were Estonian (73.5%, 1038), Russian (18.7%, 264), English (3.0%, 43), Ukrainian (2.2%, 31), Finnish (1.8%, 26), and German (0.8%, 11). Also, several more languages appeared (e.g., Italian, Armenian, Japanese).

The activities when a dictionary consultation is needed (Q20-21, answers N=5,023) included communication in Estonian (13.8%, 693 answers)⁴, producing Estonian texts (11.1%, 557), translating into Estonian (8.9%, 445), editing Estonian texts (8.7%, 436), searching for terminology (8.5%, 425), compiling (official) documents in Estonian (8.0%, 402), etc. However, not all respondents were language professionals, as might be inferred from here. We broadly categorized individuals into two groups: language professionals and general users (see Langemets et al., 2024a).

2.2 Qualitative Analysis

The respondents left a total of 1,697 additional open-ended feedback comments. For this paper, we analyzed the comments of the questions focusing on searching for information in the portal. We left out the comments connected with the publication of the paper version of the DSE 2025 (because of the updated *Action Plan*, see above Section 1) as well as comments on the layout of the portal. So, we chose four questions (in different modules) with altogether 1,049 comments:

1. Do you use the language portal *Sõnaveeb*? (243 comments, Q1)

³ The detailed description of modules and questions can be found in Langemets et al., 2024b.

⁴ We realize that this option was too vague to draw any conclusions from.

2. Are there any more linguistic issues that should be presented in the language portal *Sõnaveeb*? What information do you miss in the portal? (328 comments, Q6)
3. How satisfied are you in all with the user experience of the language portal *Sõnaveeb*? Share your comments, if you have any. (316 comments, Q3)
4. If you have any feedback about the language portal *Sõnaveeb* that you could not say previously in this questionnaire, please comment here. (162 comments, Q14)

The online tool QCAMap (Mayring & Fenzl, 2022) was used to analyze the answers and comments, enabling qualitative inductive content analysis with the support of a system of categories and rules. The inductive method helped to understand the attitudes of the participants and to explore their interpretations. We coded several topics mentioned in the comments, that can be related to the evaluative dimensions (e.g., more information, norms) as well as the general attitude of the comments (neutral, positive, negative). In the light of attitude research, the comments give a lot of data.

3. Findings from the Comments: Topics and Attitudes

Topics detected in the comments (one person could mention several topics in their comment) were as follows: seeking more information (226), users' habits of preferring other sources than *Sõnaveeb* (189), language norms (184), improvement suggestions for the portal (106), the user experience (79), reliability of the portal (54), rumors about the portal (12), bidding less information (10). Altogether 860, see Figure 1. The rest of the comments (out of 1,069) did not mention any specific topic but just reported about using the portal.⁵

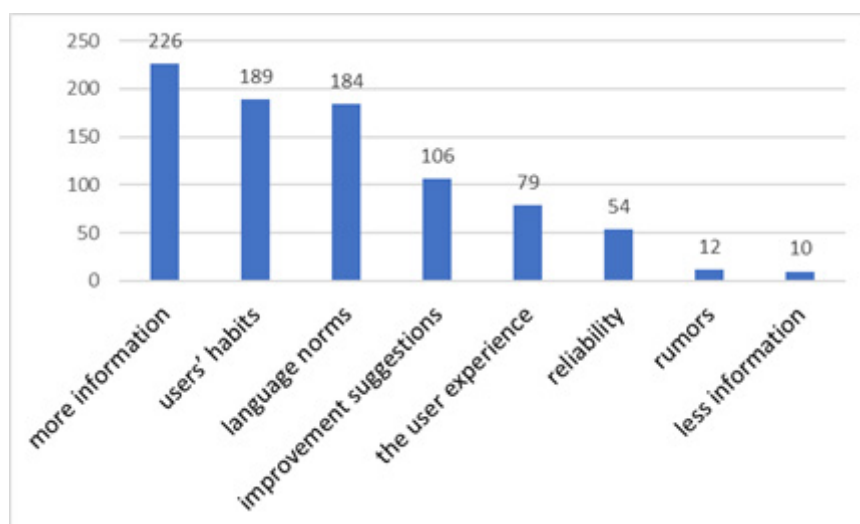


Fig. 1: Topics detected in the comments (altogether 860)

⁵ Numerical data is presented for background information to illustrate aspects that emerged from the qualitative analysis.

There were 609 comments expressing some kind of attitude: positive (207, 34.0%), negative (80, 13.1%), or both positive as well as negative attitudes (17, 2.8%). Half of the comments (305, 50.1%) stayed neutral, without expressing explicitly any attitude while commenting on topics, see Figure 2. This result correlates with the outcome of the Likert scale question (Q3, see above Section 2.1) indicating the high satisfaction (83.7%) of the survey respondents.

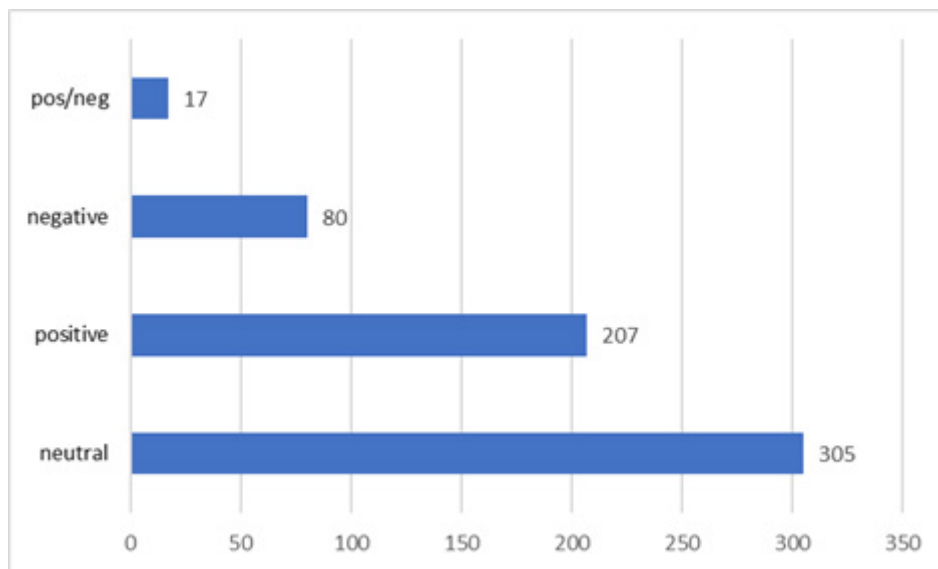


Fig. 2: Attitudes detected in the comments (altogether 609)

In the context of the ongoing discussion in society, we chose to focus on language norms (Section 3.1) and reliability (Section 3.2). There was a notable difference in attitudes: while the negative comments focused on norms (61%) or reliability (37.9%), the positive comments focused more on the content (e.g., meanings, other languages, more information) and much less on norms (23.7%) or reliability (1.9%). We specifically chose to look into what was (across all topics) commented about the content in the CombiDic – meanings (Section 3.3) and other languages (Section 3.4).

3.1 Language Norms

Language norms in Estonian primarily focus on spelling and grammar (inflection), but also on pronunciation if it affects inflection. Since 1980, it has been stated that the meanings of general language words should not be standardized; however, this understanding has not yet been firmly established in society (Risberg, 2024, p. 98). Since 2020, there have also been debates about standardizing meanings.

Since the 1990s, Estonian society has been witnessing ambivalent attitudes: while one segment of society expects commands and prohibitions, “a crusade against bad usage”, others feel that the previously predominant standard language should be regarded as just one of several equally valid sublanguages (M. Erelt, 2007, p. 333). All the norms and recommendations can more than ever be based on actual usage

due to the large corpora (nowadays Estonian National Corpus, see Koppel & Kallas, 2022). The language norms are to be found both in CombiDic (via the language portal *Sõnaveeb*) as well as the DSE, while only the latter is considered the legal source of the norms.

The norms were mentioned in 184 instances (see Figure 1). The positive comments brought up the convenient exposition of the grammar (inflection) information as the correct/appropriate forms are found easily. There's also an intuitive perception of the CombiDic as prescriptive – but it is a well-known fact that any dictionary is easily perceived that way.

The negative comments brought up issues participants were dissatisfied with. The respondents did not like the existence of two sources as they felt the need to double-check the correctness of one (CombiDic) compared to the other (DSE). We have aimed to fix this problem by integrating the lexical data from different dictionaries into one CombiDic in the portal (see above Section 1) but as a consequence of the recent legal decisions the DSE has yet to be presented separately. Participants were also worried about how they would know what is actually correct (if it is not explicitly stated) and what belongs to the literary tradition. They asked if everything had been discussed in the official Language Committee of the Mother Tongue Society. They feared that rules and norms are vanishing, if not strictly commanded by the language planners: “*DSE does not allow everything, but Sõnaveeb does*”, and expressed that it is not clear whether a word belongs to spoken or written language (although the register labels are available). The attitude of this comment is that DSE has a higher status than *Sõnaveeb*. The question of “*Who is the target audience?*” was raised as the CombiDic differs from previous DSEs (since 1999) for incorporating *all* meanings (all language data), instead of focusing only on special controversial cases.

On several occasions, different individuals expressed opposing views in the comments. For example: “*Considering the increasingly sloppy language use, Sõnaveeb is a strong pillar on which to maintain the quality of the mother tongue. You are doing commendable work!*” While just the opposite: “*The inflection tables work. [...] That's it. Otherwise, I would say that Sõnaveeb contributes to the triumph of zero style and impoverished language.*”

The need to evaluate others is a natural part of human behavior (see Cameron, 1995 on “verbal hygiene”). Some members of society exhibit “grassroots prescriptivism” – they send letters to the media about incorrect language usage (Lukač, 2018; Lukač & Heyd, 2023). In addition to critically evaluating others, some participants also criticize themselves and hesitate to express themselves, fearing ridicule from others. This is evidenced by the fact that users often seek black-and-white answers about what is ‘right’ and ‘wrong’ in language from EKI language counseling (Risberg, 2022, p. 207; Paet, 2023, p. 39, 46). This behavior can be linked to the high status of lexicographical assistance and shows traces of linguistic insecurity of language users since they feel the urge to be utterly correct.

3.2 Reliability

The reliability of the content in dictionaries is found to be important for users (Müller-Spitzer & Koplenig, 2014). The recent discussion in Estonia has somewhat turned this notion upside down when talking about relying on corpora. While Carolin Müller-Spitzer and Alexander Koplenig (2014, p. 168) found in their study that in terms of reliability, participants “considered important that all details represent actual language use and are validated on a corpus,” in Estonia, incorporating actual language usage and the use of corpus data in lexicography have caused misunderstandings and reluctance (e.g., Vider, 2022; Veldre, 2022).

Reliability was mentioned in 54 instances (see Figure 1). Positive comments reported finding it helpful to get the grammar (inflection) and meaning information easily, but a little confused if provided multiple variants: “*Which of the possible variants philologists prefer?*” It is not clear who are philologists – language planners? or copy editors? It was mentioned that more terminological data would increase the credibility of *Sõnaveeb*.

The negative comments mentioned the incompleteness, even randomness of the *Sõnaveeb*, as it is updated daily, while the previous dictionaries (DSE, explanatory, etymology, etc.) are considered well-researched and reliable. It was commented that there are “*rumors circulating that Sõnaveeb is not a recognized source*”.

In addition to the usage examples selected by the lexicographer, we present automatically selected sentences from the corpus (clickable if needed). Although there is a warning that these sentences may contain errors, they have “*given the impression that the word is widely used in standard language*”. This is another example of the influence of SLI. This is also a lesson to learn: the data should be extremely carefully arranged on the page to be better understood.

Otherwise, the negative comments were pretty much the same as mentioned in connection with the norms (see Section 3.1): users do not like to navigate between two websites (DSE and *Sõnaveeb*) and they are unsure which information is ‘correct’ and which is ‘incorrect’. Some of the respondents do not trust *Sõnaveeb* (“*Sõnaveeb is flawed*”) because it is too similar to an explanatory dictionary, while they are seeking (strict) guidelines, and language planning advice, instead. This indicates the high prestige of DSE, and that the status stereotype of language attitudes (e.g., Dragojevic et al., 2018, p. 30) is spread. This indicates that the attitudes toward *Sõnaveeb* can be perceived as reflecting a trust or distrust in authority, as opposed to a sense of solidarity. This is in contrast for example to the case of youth language, where the ‘correctness’ of language is regulated among youth who are not seeking to align themselves with any external authority (Koreinik et al., 2023).

3.3 Meanings

A study conducted in Europe has revealed that the most common queries in dictionaries are for meanings (“What does this word mean?”), as well as for spelling

and grammar (Langemets et al., 2018; Kosem et al., 2019). The same findings emerged from the *Sõnaveeb* survey (Langemets et al., 2024b), where participants were asked to rate the types of information most frequently searched: meanings were chosen the most (92.5%), followed by spelling (92.0%), then grammar (inflection) and synonyms (90.7% each). Satisfaction with language information (ratings of 4-5, meaning “Rather satisfied” and “Completely satisfied”) was as follows: spelling (88.3%), grammar (86.4%), and meanings (84.2%). The poorest satisfaction score, but still over 50%, was earned by equivalents in other languages (51.5%).

The legacy of historical language planning in Estonia has favored the “one word one meaning”-principle, which involves rejecting new senses, coining new words for different senses of polysemous words, etc. (Risberg & Langemets, 2021; Risberg, 2024). For example, the verb *vabandama* was long considered to have only the ‘correct’ meanings of ‘to forgive’ and ‘to justify,’ while the common meaning of ‘to apologize’ was deemed ‘incorrect’, as it was thought to be influenced by Russian (Raadik, 2020). Additionally, the DSE 2018 retained 130 words whose newer meanings were not recommended for use in Standard Estonian due to various language change reasons considered undesirable in language planning, such as the influence of other languages, figurative language use, and metonymy (see Risberg, 2023).

Language has for quite a long time been conceptualized by language planning as a strictly manipulable system. This has facilitated the understanding that it is possible to normatively fix word meanings in dictionaries (e.g., Aavik, 1936; Veski, 1958; T. Erelt, 2003). Such views could be cultivated in specialized terminology, but not in general language. With the emergence of the cognitive paradigm in linguistics in the 1970s and 1980s, language has increasingly been understood as part of human cognition (see Croft & Cruse, 2005). However, this knowledge has not yet reached all parts of the Estonian public awareness.

Meanings were mentioned in 78 instances (across all topics, see above Section 3). Positive (or neutral) comments favored dynamic updating, including adding new meanings, and a plentiful choice of synonyms. However, regarding polysemous words, some would like to know which meaning is the most common or preferred one. If there is a new or changed meaning (or form), some would expect an explanation and justification for the change. Some want more labels to better understand which registers the meanings belong to. Some reported improving Estonian language skills while searching through all meanings.

Negative comments, once again, revealed the longing for recommendations, and guidance on ‘correct’ and ‘incorrect’ meanings: “*Not every new meaning is correct.*” As previously described, we register meanings emerging from actual language use in CombiDic, rather than capturing any arbitrary usage instances. However, the Standard Language Ideology comes into play here, leading some members of society to believe that word meanings should not change (see also Trudgill, 1998).

3.4 Other Languages

In the language portal *Sõnaveeb*, the default option is to present all lexical data we have assembled in our database, including all languages, only then to be further filtered. Russian is the best-represented other language in the CombiDic,⁶ as we have imported the voluminous Estonian-Russian (1997–2009, 2nd ed. 2019) dictionary into our database Ekilex (Tavast et al., 2018, Koppel et al., 2019). French and Ukrainian are represented in the range of the basic vocabulary (ca 6,000 words), while English is only random, and German is missing completely (so far, August 2024).

Other languages were mentioned in 101 instances (across all topics, see above Section 3). Attitudes towards other languages are roughly divided into two categories: some would like to see even more equivalents in different languages, while others do not want any other languages to be seen in the Estonian dictionary.

Positive comments approved of the presence of other languages: “*Since Sõnaveeb has started adding foreign language equivalents as well, it should confidently continue in that direction.*” Neutral comments proposed adding more equivalents in more languages, or the opposite, stated that they are not seeking equivalents in other languages. It was also noted that some users might prefer separate websites, meaning they would like to use the Estonian monolingual dictionary separately from other languages.

The lack of English equivalents was mentioned most frequently. Some specified that they would like to see more (official) European languages. German and Finnish were also mentioned more, less so Latvian, Ukrainian, French, and Swedish. Some respondents would like to see even more Russian equivalents.

As for negative comments, respondents feared it to be misleading containing French and English in the CombiDic, as the users could make mistakes in Estonian spelling when accidentally choosing the form of the foreign word. Some of the negative comments were extreme: “*Please remove the blocks of foreign words/definitions, they are unnecessary noise*” or “*Sõnaveeb is a nightmare. Who is it meant for? Russians?*”

The negative attitude towards Russian can be explained with historical reasons, as Estonia had been occupied by Russia for half a century (1940–1991): “*Instead of Russian translations, there could be English ones*” or “*Why does every word have a Russian equivalent?*”

4. Discussion

The rather poetic question in the title of this paper asked if one should dream or not dream about ‘correct’ meanings. Depending on the viewpoint, the answer can be

⁶ There are more than 140 terminological databases in the *Sõnaveeb* that include equivalents mostly in English but also many more languages.

different. Although only a handful of respondents expressed dissatisfaction with the language portal *Sõnaveeb* (4.7%, 80 out of 1,697 comments, see Figure 2), the open-ended comments revealed various concerns and attitudes. In general, attitudes were divided based on whether the user was seeking information to enrich their language or seeking language norms and language planning recommendations. Meanings were not an exception: there were users who definitely believe in 'correct' meanings guarded by lexicographers or language planners and others who expected neutral descriptions of all meanings in use, and perhaps expected some labels for registers.

The study reveals that the online survey participants' attitudes towards dictionaries (lexicographical assistance) and the portal *Sõnaveeb* can be placed on the scale of status (Dragojevic et al., 2018, p. 30). Status is linked to the idea that *Sõnaveeb* is for the majority of users reliable and thus carries a high status. Another similar attitude is that it has prescriptive power (norms) and thus has high prestige. Dictionaries are universally seen as a standard that can be referred to as an authority (e.g., Müller-Spitzer, 2014, p. 85), and as language codifiers that have a social force to determine what is standard in a language (Ammon, 2015, p. 57) even more. It is unsurprising that they have a status-based stereotype rather than a solidarity-based one. The latter would mean that participants do not seek authority from above; instead, language use would be regulated among users. A good example is youth language, where solidarity between youth sets the standards for correctness (Koreinik et al., 2023).

The normative and descriptive approaches have always collided in lexicography⁷ where in the case of contemporary Estonian, the language norms are provided for spelling (orthography) and grammar (inflection), but not for meanings (in general language). The same has been stated for Danish: "The usual way for lexicographers to deal with meanings is to include them (descriptively) in the dictionary but supply them (normatively) with a label that indicates the nature of the restriction on their usage." (Trap-Jensen, 2002) So, labels can function a bit like normative advice, but of course, only to some extent. Labels should be perceived as tendencies, not the universal norm across language, as each language situation depends on the context.

The Finnish linguist Pirkko Nuolijärvi (2023) has argued regarding the Estonian DSE that "If the dictionary is titled *Õigekeelsussõnaraamat* [= DSE, literally 'Correct Language Dictionary'], then the user of the dictionary easily interprets all recommendations as norms: this word form or meaning is correct." The title might have 'worked' in the period of codifying the standard language (i.e., in the 1920s for modernizing Estonian), but since the language has stabilized at least since the 1980s, it seems anachronistic (see also Risberg 2024; Kerge 2012). The latest developments in Estonia have regrettably proved the opposite (see "Press release, 28.03.2024"): we should still separately publish the normative DSE (validated by the law), while the option would have been to proceed with one source, the descriptive *Dictionary of Estonian*, which would have had some usage notes by the DSE in a separate column.

⁷ As well as within society more broadly – some members do long for normative guidelines for the entire language.

From the CAAE interviews, it became clear that language professionals understand language more flexibly than general users, who strive to be correct in both formal and informal writing (Langemets et al., 2024a). But even L1 language professionals seek greater assurance from the DSE regarding the appropriateness of a word or its meaning for Standard Estonian. This attitude has been fostered by the long tradition of two types of general dictionaries of Estonian: the normative DSEs (since 1918) and descriptive explanatory dictionaries (since 1988). The common perception of language in Estonia has long been influenced by Standard Language Ideology, which holds the belief that one language variant (in our case Standard Estonian) is superior to others, and purism, i.e., the avoidance of undesirable language elements and foreign influences (Lindström, Risberg, & Plado, 2023; Paet, 2023).

In the online survey, it was notable that the few (80) negative comments focused on norms and reliability, while respondents with a positive (207) or a neutral (305) attitude commented much more on many other topics (e.g., other languages, meanings, layout). It might be connected with the belief in the ‘correct’ language, or be the simplest way to criticize the dictionary one does not favor. In Estonia, relying on corpora has caused mistrust in the public as well, although the corpora have been used since the 1990s, but not for compiling the normative DSE. These kinds of negative comments have been widely reflected in the media during the public discussion in recent years. A behavioral scientist has argued that conflicting contents are usually produced by a minority: “About one-fifth or less produce four-fifths or more of such content, including competitive content (Uusberg, 2024).”

Descriptive approach is considered a cornerstone of scientific linguistic research, emphasizing the importance of factual accuracy. In contrast, prescriptivism lacks a scientific nature – the firm stances taken by prescriptivists do not necessarily reflect actual language usage and thus fall short of scientific standards. (Ayres-Bennett, 2021, p. 44) Consequently, a primary goal of Estonian modern language planning, based on the results of research conducted within the framework of usage-based linguistics, is to harmonize language norms and recommendations with the inherent norms of actual language usage, minimizing potential conflicts. When standards closely reflect the natural patterns of language, users are less reliant on rote memorization, resulting in more efficient language use (Risberg, 2024, p. 100; Päll, 2019). The descriptive approach has already been applied for compiling and editing the CombiDic, and partly, to the extent possible in a language planning dictionary, it will also be used for the DSE 2025.

5. Conclusion

In 2023, the Institute of the Estonian Language conducted a user experience survey to better understand the users of the language portal *Sõnaveeb* and their needs and concerns. Since 2019, the portal has been developed with the goal of consolidating all lexical data into one place, integrating general dictionaries into the *EKI Combined Dictionary* (CombiDic). Although the survey did not specifically ask about users’ attitudes, these were evident in the open-ended comments. We found that the portal

users fell into three groups based on their opinions and attitudes about language: (a) users seeking to find increasing amounts of information, who left positive comments; (b) users looking to enrich their language, who left both positive and neutral comments; and (c) users primarily seeking correctness and guidelines, who left negative comments.

The range of normative aspects of language – spelling and grammar (inflection), and pronunciation if it affects inflection – should be more clearly explained in Estonian society, as this understanding has not yet been firmly established. Regarding meanings, recent research (Risberg, 2024; Paet, 2023) has once again proven that one should not dream of 'correct' meanings but trust the meanings described in the dictionary.

References

- Aavik, J. (Ed.). (1936). *Eesti õigekeelsuse õpik ja grammatika*. Noor-Eesti.
- Action Plan 2023 (renewed in 2024) = *ÕSi ja EKI teatmiku koostamise tegevuskava 2022-2025*. Retrieved August 1, 2024, from <https://eki.ee/wp-content/uploads/2024/05/os-ja-eki-teatmiku-koostamise-tegevuskava.pdf>
- Ammon, U. (2015). On the social forces that determine what is standard in a language – with a look at the norms of non-standard language varieties. *Bulletin suisse de linguistique appliquée. Numéro spécial*, 3, 53–67.
- Cameron, D. (1995). *Verbal Hygiene*. Routledge.
- CombiDic = Langemets, M., Hein, I., Jürviste, M., Kallas, J., Kiisla, O., Koppel, K., Kuusk, K., Leemets, T., Mäearu, S., Paet, T., Päll, P., Raadik, M., Risberg, L., Rehema, T., Tammik, H., Tiits, M., Tsepelina, K., Tuulik, M., Uibo, U., Valdre, T., Viks, Ü., Dovgan, E., Sai, E., Smirnova, A., Tubin, V., & Ossiptšuk, I. (Eds.). (2024). *Eesti Keele Instituudi ühendsõnastik* ['EKI Combined Dictionary']. Eesti Keele Instituut, Sõnaveeb ('Word Web'). Retrieved August 1, 2024, from <https://sonaveeb.ee>
- Croft, W., & Cruse, D. A. (2005). *Cognitive Linguistics*. 3. ed. Cambridge University Press.
- Díaz-Campos, M., & Balasch, S. (Eds.). (2023). *The Handbook of Usage-Based Linguistics*. Wiley-Blackwell.
- Diessel, H. (2017). Usage-based linguistics. In M. Aronoff (Ed.), *Oxford Research Encyclopedia of Linguistics*. Oxford University Press.
- Dragojevic, M., Fasoli, F., Cramer, J., & Rakić, T. (2021). Toward a Century of Language Attitudes Research: Looking Back and Moving Forward. *Journal of Language and Social Psychology*, 40(1), 60–79. <https://doi.org/10.1177/0261927X20966714>
- Dragojevic, M., Berglund, C., & Blauvelt, T. K. (2018). Figuring Out Who's Who: The Role of Social Categorization in the Language Attitudes Process. *Journal of Language and Social Psychology*, 37(1), 28–50. <https://doi.org/10.1177/0261927X17706942>

Dziemianko, A. (2012). On the use(fulness) of paper and electronic dictionaries. In S. Granger, & M. Paquot (Eds.), *Electronic Lexicography* (pp. 319–342). Oxford University Press.

Ebner, C. (2017). *Proper English Usage: A sociolinguistic investigation of attitudes towards usage problems in British English*. [Doctoral dissertation, University of Leiden].

Edwards, J. R. (1999). Refining our understanding of language attitudes. *Journal of Language and Social Psychology*, 18, 101–110.

EKSS = Langemets, M., Tiits, M., Valdre, T., Veskis, L., Viks, Ü., & Voll, P. (Eds.). (2009). *Eesti keele seletav sõnaraamat*. Vol I–VI. Eesti Keele Sihtasutus.

Erelt, M. (Ed.) (2007). Estonian language. *Linguistica Uralica supplementary series*, 1. Eesti Teaduste Akadeemia Kirjastus.

Erelt, T. (2002). *Eesti keelekorraldus*. Eesti Keele Sihtasutus.

Erelt, T. (2003). ÕS-i toimetaja selgitab. *Keel ja kirjandus*, 6, 448–460.

Fazio, R. H., & Olson, M. A. (2003). Attitudes: Foundations, functions, and consequences. In M. A. Hogg, & J. Cooper (Eds.), *The Sage Handbook of Social Psychology* (pp. 139–160). Sage.

Garrett, P. (2010). *Attitudes to Language*. (Key Topics in Sociolinguistics.) Cambridge University Press.

Giles, H., & Marlow, M. (2011). Theorizing Language Attitudes Existing Frameworks, an Integrative Model, and New Directions. *Annals of the International Communication Association*, 35(1), 161–197. <https://doi.org/10.1080/23808985.2011.11679116>

Keeleminutid 13.06.2023 = Sõnaveebi kasutajauuringu tööühm (Eds.) (2023). Keeleminutid. Suvi pole keeletu: keenialane New Yorgis. *ERR Kultuur*, 13.06. Retrieved August 1, 2024, from <https://kultuur.err.ee/1609005533/keeleminutid-suvi-pole-keeletu-keenialane-new-yorgis>

Kerge, K. (2012). Keelekorralduse põhimõtted. *Sirp*, 27.09. Retrieved August 1, 2024, from <https://www.sirp.ee/s1-artiklid/varamu/keelekorralduse-pohimotted/>

Koppel, K., Tavast, A., Langemets, M., & Kallas, J. (2019). Aggregating dictionaries into the language portal Sõnaveeb: issues with and without a solution. In I. Kosem, T. Zingano Kuhn, M. Correia, J. P. Ferreira, M. Jansen, I. Pereira, J. Kallas, M. Jakubiček, S. Krek, S., & C. Tiberius (Eds.), *Proceedings of the eLex 2019 conference. 1-3 October 2019, Sintra, Portugal* (pp. 434–452). Lexical Computing CZ, s.r.o.

Koreinik, K. (2023). Extra-linguistic arguments in the XXI century corpus and status planning: Superdictionary in times of language standardization from below. *Eesti ja Soome-Ugri Keeleteaduse Ajakiri. Journal of Estonian and Finno-Ugric Linguistics*, 14(1), 117–148. <https://doi.org/10.12697/jeful.2023.14.1.04>

Koreinik, K., Mandel, A., Pilvik, M.-L., Praakli, K., & Vihman, V.-A. (2023). Outsourcing teenage language: a participatory approach for exploring speech and text messaging. *Linguistics Vanguard*, 9(4), 389–398. <https://doi.org/10.1515/lingvan-2021-0152>

- Kosem, I., Wolfer, S., Lew, R., & Müller-Spitzer, C. (2018). Attitudes of Slovenian Language Users Towards General Monolingual Dictionaries: an International Perspective. *Slovenščina* 2.0, 6(1), 90–134. <http://dx.doi.org/10.4312/slo2.0.2018.1.90-134>
- Kosem, I., Lew, R., Müller-Spitzer, C., Ribeiro Silveira, M., Wolfer, S., & 54 more (2019). The Image of the Monolingual Dictionary Across Europe. Results of the European Survey of Dictionary use and Culture. *International Journal of Lexicography*, 32(1), 92–114. <https://doi.org/10.1093/ijl/ocy022>
- Langemets, M., Uibo, U., Tiits, M., Valdre, T., & Voll, P. (2018). Eesti keel uues kuues. Eesti keele sõnaraamat 2018. *Keel ja Kirjandus*, 12, 942–958. <https://doi.org/10.54013/kk733a2>
- Langemets, M., Risberg, L., Paet, T., & Algvere, K. (2024a). Keeleminutid. Kas Sõnaveebi kasutajad eelistavad korrektsust või mängulust? *ERR Kultuur*, 05.02. Retrieved August 1, 2024, from <https://kultuur.err.ee/1609243326/keeleminutid-kas-sonaveebi-kasutajad-eelistavad-korreksust-voimangulusti>
- Langemets, M., Risberg, L., Paet, T., & Algvere, K. (2024b). *Sõnaveebi kasutajauuring 2023. Raport*. ('The user experience survey 2023 of Sõnaveeb. The report'.) Eesti Keele Instituut. Retrieved August 1, 2024, from https://eki.ee/wp-content/uploads/2024/03/sonaveebi-kasutajauuring_2023.pdf
- Language Act 2011 = RT I, 18.03.2011, 1. Riigikogu ('Parliament'). Retrieved August 1, 2024, from <https://www.riigiteataja.ee/en/eli/ee/Riigikogu/act/528072023001/consolide>
- Lew, R. (2015a). Opportunities and limitations of user studies. *OPAL - Onlinepublizierte Arbeiten zur Linguistik*, 2, 6–16.
- Lew, R. (2015b). Research into the use of online dictionaries. *International Journal of Lexicography*, 28(2), 232–253.
- LimeSurvey, LimeSurvey Project Team. Retrieved August 1, 2024, from <https://www.limesurvey.org/>
- Lindström, L., Risberg, L., & Plado, H. (2023). Language ideologies and beliefs about language in Estonia and Estonian language planning. *Eesti ja Soome-Ugri Keeleteaduse Ajakiri. Journal of Estonian and Finno-Ugric Linguistics*, 14(1), 7–48. <https://doi.org/10.12697/jeful.2023.14.1.01>
- Lukač, M. (2018). Grassroots Prescriptivism: An analysis of individual speakers' efforts in maintaining the standard language ideology. *English Today*, 34(4), 5–12. <https://doi.org/10.1017/S0266078418000342>
- Lukač, M., & Heyd, T. (2023). Grassroots prescriptivism. In J. C. Beal, M. Lukač, & R. Straaijer (Eds.). *Routledge Handbook of Prescriptivism*. Routledge. <https://doi.org/10.4324/9781003095125-16>
- Mayring, P., & Fenzl, T. (2022). *QCAmap: a software for Qualitative Content Analysis*. Retrieved August 1, 2024, from <https://www.qcmap.org/ui/en/home>
- Milroy, J. (2001). Language Ideologies and the Consequences of Standardization. *Journal of Sociolinguistics*, 5(4), 530–555. <https://doi.org/10.1111/1467-9481.00163>
- Müller-Spitzer, C. (Ed.). (2014). *Using Online Dictionaries*. (Lexicographica: Series Maior 145.) De Gruyter.

Müller-Spitzer, C., & Koplenig, A. (2014). Online dictionaries: expectations and demands. In C. Müller-Spitzer (Ed.), *Using Online Dictionaries*. (Lexicographica. Series Maior 145.) (pp. 143–188). Walter de Gruyter.

Nuolijärvi, P. (2023). Võõrsõnade esitamine eesti keele sõnastikes. *Keel ja Kirjandus*, 12, 1218–1221. Retrieved August 1, 2024, from <https://keeljakirjandus.ee/wp-content/uploads/2023/12/P.NUOLIJARVI.pdf>

Paet, T. (2023). *Võõrainese kinnistumine eesti keeles: keelekorralduslik ja leksikograafiline vaade* [Doctoral dissertation, University of Tartu]. Tartu Ülikooli Kirjastus.

Press release 28.03.2024 = *Järgmise ÕSi väljaandmiseks muudetakse tegevuskava*. Haridus- ja Teadusministeerium. Retrieved August 1, 2024, from <https://www.hm.ee/uudised/jargmise-osi-valjaandmiseks-muudetakse-tegevuskava>

Raadik, L. (2020). Sõnatähenduste normimisest eesti keelekorralduses verbi *vabandama* näitel. *Keel ja Kirjandus*, 10, 853–874. <https://doi.org/10.54013/kk755a3>

Regulation 2011 = RTI, 14.06.2011, 3. *Rules for the application of the norms of the Estonian standard language*. Retrieved August 1, 2024, from <https://www.riigiteataja.ee/akt/114062011003>

Risberg, L. (2022). Mis on peidus sõnaraamatu tähendussoovituste tagahoovis? Eesti omasõnade käsitus. *Eesti ja soome-ugri keeleteaduse ajakiri. Journal of Estonian and Finno-Ugric Linguistics*, 13(2), 185–214. <https://doi.org/10.12697/jeful.2022.13.2.06>

Risberg, L. (2023). Recommendations for the meanings of words by Estonian language planning – justified and necessary, or not? *Taikomoji kalbotyra*, 20, 53–73. <https://doi.org/10.15388/Taikalbot.2023.20.5>

Risberg, L. (2024). Paronüümide kosmos. Eesti ja saksa keelt kõrvutav käsitus. *Keel ja Kirjandus*, 6, 570–588. <https://doi.org/10.54013/kk798a4>

Risberg, L., & Langemets, M. (2021). Paronüümide probleem eesti keeles. *Keel ja Kirjandus*, 10, 903–926. <https://doi.org/10.54013/kk766a4>

Tauli, V. 1943. J. V. Veski eesti keele arendajana. *Postimees*, 143, 4.

Tavast, A., Langemets, M., Kallas, J., & Koppel, K. (2018). Unified Data Modelling for Presenting Lexical Data: The Case of EKILEX. In J. Čibej, V. Gorjanc, I. Kosem, & S. Krek (Eds.), *Proceedings of the XVIII EURALEX International Congress Lexicography in Global Contexts 17-21 July 2018, Ljubljana* (pp. 749–761). Znanstvena založba Filozofske fakultete Univerze v Ljubljani.

Trap-Jensen, L. (2002). Descriptive and Normative Aspects of Lexicographic Decision-Making: The Borderline Cases. In R. Vatvedt Fjeld, & J. M. Torjusen (Eds.), *Proceedings of the 15th EURALEX International Congress. 7-11 August 2012. Oslo* (pp. 425–432). Department of Linguistics and Scandinavian Studies, University of Oslo.

Trudgill, P. (1998). The Meanings of Words Should Not be Allowed to Vary or Change. In L. Bauer, & P. Trudgill (Eds.), *Language Myths* (pp. 1–8). Penguin Books.

Uusberg, A. (2024). Inimesed otsivad konfliktist teisi psühholoogilisi hüvesid. Intervjueeris Merilin Pärli. *ERR Novaator*, 09.03. Retrieved August 1, 2024, from <https://novaator.err.ee/1609276931/andero-uusberg-inimesed-otsivad-konfliktist-teisi-psuhholoogilisi-huvesid>

Vaicekauskienė, L. (2012). 'Good Language' and Insecure Speakers: A Study into Metalinguistic Awareness of TV and Radio Journalists in the Context of Language. A. Usonienė, N. Nau, & I. Dabašinskienė (Eds.), *Multiple Perspectives in Linguistic Research on Baltic Languages* (pp. 76–103). Cambridge Scholars Publishing.

Vainik, E., & Paulsen, G. (2023). Kujundlik mõtlemine 2020.–2022. aasta keelekriisis. *Keel ja Kirjandus* 7, 651–677. <https://doi.org/10.54013/kk787a1>

Veldre, A. (2022). Kust on pärit zuumeri eesti keel? *Sirp*, 02.09. Retrieved August 1, 2024, from <https://www.sirp.ee/s1-artiklid/varamu/kust-on-parit-zuumeri-eesti-keel/>

Veski, J. V. (1958). *Johannes Voldemar Veski keelelisi töid*. Eesti Riiklik Kirjastus.

Vider, K. (2022). Sõna vabadusest, andmepõhiselt. *Sirp*, 17.06. Retrieved August 1, 2024, from <https://www.sirp.ee/s1-artiklid/varamu/sona-vabadusest-andmepohiselt/>

Wolfer, S., Nied Curcio, M., Silva Dias, I., M., Müller-Spitzer, C., & Jose Dominguez Vazquez, M. (2018). Combining Quantitative and Qualitative Methods in a Study on Dictionary Use. In J. Čibej, V. Gorjanc, I. Kosem, & S. Krek (Eds.), *Proceedings of the XVIII EURALEX International Congress Lexicography in Global Contexts 17-21 July 2018, Ljubljana* (pp. 101–112). Znanstvena založba Filozofske fakultete Univerze v Ljubljani.

Wolfer, S., Michaelis, F., & Müller-Spitzer, C. (2021). Nutzungsstatistiken von linguistischen Online-Ressourcen: Potentiale und Limitationen. *Zeitschrift für Angewandte Linguistik*, 75, 395–430. <https://doi.org/10.1515/zfal-2021-2070>

Acknowledgements

We would like to thank R&D project “Applying large language models to lexicography: new opportunities and challenges” (EKKD-III1, Institute of the Estonian Language, principal investigator: Maria Tuulik) and R&D project “Morphosyntactic variation in Estonian” (EKKD-TA2, University of Tartu, principal investigator: Liina Lindström). We would also like to thank Sascha Wolfer (Leibniz-Institut für Deutsche Sprache) for his helpful comments on how to conduct an online survey, Tiina Paet (Institute of the Estonian Language) for her contribution to the compilation and analysis of the online survey, and Marja Vaba for her insightful input at the beginning of the project.

Contact information

Margit Langemets

Institute of the Estonian Language
margit.langemets@eki.ee

Lydia Risberg

Institute of the Estonian Language

University of Tartu

lydia.risberg@eki.ee

Kristel Algvere

University of Tartu

kristel.algvere@ut.ee

Tinatin Margalitadze and Katalin P. Márkus

CROSS-BORDER COLLABORATION IN TEACHING DICTIONARY SKILLS

Abstract The paper presents a project devised by Georgian and Hungarian lexicographers which aims at improving dictionary use skills and dictionary culture in Georgia and Hungary. The project is based on previous experience, studies and findings of its authors at Ilia State University (Georgia) and Károli Gáspár University of the Reformed Church in Hungary. The feedback gathered from theoretical lexicography courses and the needs of the students emerging from these courses revealed the necessity to concentrate more on practical issues of teaching dictionary skills. The cross-border cooperation project will be divided into two stages. At the initial stage, the shortcomings in dictionary skills among students, as well as the special needs of students and teachers will be identified with the help of a questionnaire, supplemented by interviews and tests to refine the data. The results of the survey will be used in the next stage for the development of teaching materials, which will include a workbook (in print and e-book format), a teacher's book (in print and e-book format), as well as a variety of online tools and exercises for language learners which will help them explore reference skills from many different angles and in different situations.

Keywords dictionary didactics; dictionary use skills; manual for teachers; exercise book; practical lexicography module

1. Introduction

This paper presents a project, conceived by Georgian and Hungarian lexicographers at Ilia State University (ISU) and Károli Gáspár University of the Reformed Church in Hungary (KRE), aiming at developing dictionary use skills in university students. The project is motivated by our practical experience as well as our prior research projects. Many studies were conducted in Georgia concerning dictionary use. Some of them were more general aiming at revealing the attitude of the Georgian population towards dictionaries. Other studies were more specific and investigated the effectiveness of different types of dictionaries – monolingual, bilingual or bilingualized – in English language classes. Particularly interesting are those experiments which tried to measure the dictionary use skills of Georgian learners of English. The above-mentioned studies and the obtained results are described in detail in the article “How to Solve Problems in Dictionary Use: the Georgian Case” (Margalitadze & Meladze, 2023). Research projects at KRE were focused on dictionary didactics, dictionary use and how dictionaries could best be employed in the context of language teaching and learning (for the results of an ongoing research project, see P. Márkus et al., 2023). Research on dictionary use is an essential prerequisite for high-quality dictionaries and effective dictionary use. In addition, it has the potential to improve the teaching of dictionary use and lexicographic culture. For this reason, we plan to launch a joint research project to explore dictionary use habits and attitudes of university students

learning English as a foreign language (EFL) in Georgia and Hungary, as well as to identify the current obstacles to effective dictionary use.

The studies carried out in Georgia and Hungary as well as the previous dictionary user studies (e.g., Rundell, 1999; Lew, 2010; Chan, 2012; Neid Curcio, 2022; Gouws, 2022; P. Márkus & Dringó-Horváth, 2023; Vrbinc & Vrbinc, 2006) have demonstrated that there is a significant gap between the dictionary user and the dictionary since people lack the reference skills required to retrieve information in a complex dictionary entry. Moreover, not having studied dictionary use in an educational context, practising teachers do not know how to integrate the topic into classroom activities effectively. There could be two approaches to resolving the dichotomy between the user and the dictionary. First, lexicographers are trying to develop techniques that match ordinary users' linguistic expertise and make dictionaries more user-oriented. As an example, online platforms are increasingly enabling this, as digital dictionaries have a much more flexible structure and fewer structural constraints than their paper predecessors. E-dictionaries are products that are ideally designed to suit human needs; thus, it seems reasonable that they would be user-friendly in every way, including editorial layout and design (cf. Tarp & Gouws, 2020). Second, users should be trained to be more skilled in using dictionaries and other lexicographic tools because no matter how user-friendly the source is, in order to successfully retrieve data, the user needs to know exactly what kind of data set is represented in different dictionary types. Knowing the layout and the basic properties of data structure is essential to accessing the right information (cf. Margalitadze & Meladze, 2023; P. Márkus & Dringó-Horváth, 2023). Students need to know the internal structure of the textual content (e.g., of an entry) that they read in print or on the internet, otherwise, the retrieval of information will take a lot of time or fail completely (cf. Lew, 2013; 2016; Campoy-Cubillo, 2015; Tarp & Gouws, 2020). In reality, we witness the improvement in the quality and user-friendliness of dictionaries, however, as a result of the lack of dictionary skills training at schools and universities, societies are losing dictionary culture (cf. Gavriilidou & Konstantinidou, 2022; Gouws, 2022). The consequence of this is poor knowledge of foreign languages, scanty vocabularies, and the negative effect it entails for national languages (Margalitadze, 2018; 2020).

There is one more aspect of dictionary use skills. In today's internet-driven society, an increasing number of students are studying foreign languages outside of the traditional classroom, which may provide challenges in finding reliable sources of information (cf. Nied Curcio, 2022; P. Márkus et al., 2023). In general, there is a considerable difference between classroom learning and autonomous learning. Since language teachers are not always available, students who are having language-related problems must look for credible sources of information on their own. If students learn how to utilize reference works, such as dictionaries or encyclopaedias efficiently, these resources will help them to answer their language-related questions. Thus, they will make a big step in the direction of becoming autonomous learners of English (cf. Roby, 2006; Lew, 2016; Rundell, 2016).

Based on our previous experience and findings, we embark on jointly developing a dictionary skills handbook that can be integrated into lessons in different courses,

at different language levels, or even for independent study. The handbook will also include a methodological guide for teachers to give them ideas and useful advice on how to teach dictionary skills.

2. Lexicography Courses at ISU and KRE

A course in lexicography, “Introduction to Lexicography” (Margalitadze, 2022), was introduced at ISU and several state and private universities in Georgia. The course consists of two parts: a theoretical part, dealing with different issues of modern lexicography and a practical part, made up of exercises developed by the author of the course. The exercises are tailored to the needs of Georgian students and address the most problematic issues, mostly connected to the meaning of words, understanding the complexity of interaction between English and Georgian lexis, the problem of equivalence, polysemy, homonymy, conversion, phrasal verbs and idioms. Students’ feedback is particularly positive on the exercises which they find very useful and efficient in improving their dictionary use skills. Students not only improve their dictionary skills but also understand the importance of dictionaries in learning foreign languages, in enriching their vocabulary, in becoming more independent in learning foreign languages.

At KRE, “Lexicology and lexicography”, as university courses, have been taught since the academic year 1996/1997 (at the Department of English Linguistics). The lexicology part of the course provides a basis for many of the lexicographic topics covered in the second section. The objective is to present the theoretical and practical relationships between these two important fields and to offer all the resources required for hands-on dictionary skills training. Research shows (Márkus & Szöllősy, 2006; P. Márkus et al., 2023) that these courses should increasingly focus on practical lexicography, covering areas such as dictionary use and reference skills, as this will be an important part of the digital literacy of students (whether linguists or teachers) graduating from university.

The lack of available textbooks and teaching aids is tangible in this area. Studies reveal that in Hungary and Georgia “dictionary awareness” is generally rather low and that more attention to the teaching of dictionary skills would be needed in the curricula for English language learning. The major challenges that result from a weak dictionary culture, as well as teachers’ inability to incorporate dictionary pedagogy into regular teaching activities in primary and secondary schools, are readily obvious. The following section will detail our cooperative project, which aims to find ways to overcome some of these problems.

3. New Project

Our new project “Cross-Border Collaboration in Teaching Dictionary Skills” advocates for the inclusion of dictionary skills in language classes at school and university. Through our university lexicographic courses, more and more teachers are recognising the importance of teaching and integrating the use of dictionaries

into the educational process. After graduation, students give feedback on how they continue to incorporate dictionary use into foreign language teaching. In order to facilitate these positive trends, we must constantly strive to provide methodological aids and training courses to help language teachers achieve a more diverse and cross-competence development – using a wider range of digital tools (e.g., ChatGPT) and adapting to the changing needs of students.

Our cross-border collaboration aims to demonstrate the teachability of dictionary skills, and to this end, we attempt to present the topic from multiple perspectives, involving as many activities as possible. The first thing teachers need to learn is that knowledge is not simply a matter of repeating existing knowledge, but rather of acquiring the missing knowledge as effectively as possible, for which creative thinking and problem-solving are essential skills (Baron, 2021). Moreover, it is crucial to understand and distinguish between two fundamental forms of dictionary use: active and passive. We aim to develop a conscious use of dictionaries in learners, hence exercises will be developed for both types of dictionary use. Dictionary skills exercises will focus on all the information types (e.g., phonological, grammatical) in the dictionary entry (both monolingual and bilingual). From a methodological point of view, there are several key aspects to keep in mind concerning the tasks, such as *recognition* (e.g., exercises focusing on recognition of the correct part of speech, correct homonym, correct polysemous meaning of a word); *comparison* (e.g., students compare a dictionary entry from an online dictionary with a print dictionary); *evaluation* (e.g., comparing dictionaries according to evaluation criteria); *observation* (e.g., tasks interpreting the structure of a dictionary entry or the user guide require observational activity). This type of task is particularly important since careful observation is required for solid thinking. Critical thinkers collect data during observations and filter out extraneous aspects to focus on the important facts, which is considered an essential skill in the 21st-century digital world. Dictionaries are an inexhaustible repository for grouping words in different ways, so they can be used for a variety of topics, as the language learner can search for words of the same part of speech, the same ending, or the same meaning in print or electronic dictionaries. Exercises involving the *collection* of words are suitable for teaching the effective use of both print and electronic dictionaries. The *supplementation* activity can also be carried out creatively, for example, in online dictionaries, where the task is to further edit or supplement individual entries. Finally, *creativity* and *cooperation* should also be highlighted (e.g., students prepare their own tasks based on the given sample exercises independently or in groups) (Gonda, 2009; P. Márkus, 2023).

Some of our experiments revealed that students try to use ChatGPT, Google Translate or other tools instead of dictionaries while completing different vocabulary tasks (cf. P. Márkus et al., 2023). Therefore it is important to include these tools for the development of dictionary use skills. It is important to teach learners of English to what extent they can rely on modern tools and what their pitfalls and limitations are, particularly in a bilingual context while translating English texts into their native languages or vice versa. Such comparisons between the performance of machine translation or ChatGPT and a dictionary entry will prevent students from being too dependent on technologies and will help them realize that these tools should be

regarded as aids in solving particular tasks and not the only resources to be consulted. In order to develop a method to improve the existing situation in dictionary didactics, this project intends to investigate where and when the instructors of the different courses educate students on the skills of dictionary use in university programmes. In light of all this before embarking on developing a dictionary skills module the current situation will be assessed by identifying:

- the shortcomings of dictionary skills among the 1st-year students;
- the special needs of students;
- the special needs of teachers.

For this purpose, a questionnaire (which aims to investigate students' dictionary use habits and attitudes towards dictionaries, and the role of dictionaries as an aid to language learning and teaching; for more on the questionnaire, see P. Márkus et al., 2023), devised by Hungarian lexicographers, will be used both at ISU and KRE. The questionnaire will be supplemented in some cases (approximately 10%) by interviews. The aim of the interviews is to refine the quantitative data and to highlight further problems and needs. The results of the survey at both universities will be thoroughly analysed by the project participants. Based on the findings, a test-based experiment will be conducted at ISU and KRE. The experiment will start with a pre-test, worked out at ISU, for measuring the dictionary use skills of students. Based on the pre-test, we will develop worksheets aimed at effective dictionary use, focusing on problem areas. During one semester these teaching materials (i.e., worksheets) will be tested in the classroom in various courses. Students will be asked for feedback on the materials, as we want to know what they think about the tasks, and how difficult or useful they find them. At the end of the semester, the "dictionary skills training" will end with a post-test to measure students' progress and see how their skills have developed. The feedback and the results of the post-test will be used to develop a "dictionary skills module" based on the worksheets that will be constantly improved and revised during the semester. Thus two clearly distinctive but interrelated parts will be dealt with during the project. First, dictionary use habits, strategies and needs will be mapped by the questionnaire, interviews and tests, identifying areas for improvement. Second, a "dictionary skills module" intended to improve these skills will be designed and tested.

As a result of the project, we will jointly develop a dictionary skills handbook with exercises and activities that can be integrated into lessons in different courses and language levels. The handbook will also include methodological guidelines on dictionary didactics for teachers to give them ideas and useful advice on how to teach dictionary skills. The teaching materials will include a workbook (in print and e-book format), a teacher's book (in print and e-book format), as well as a variety of online tools and exercises so that language learners will be able to explore reference skills from many different angles and in different situations.

4. Conclusion

We believe that the successful implementation of the project and development of the practical lexicography module will contribute to the improvement of dictionary culture in our countries. We are also convinced that good dictionary skills will enable students to be more successful in their English degree programmes – and not only in the language development classes, but also in the literature, history, culture, and linguistics courses. Improving students' efficiency in using online dictionaries of English will also improve their digital competencies in general, and data mining and data retrieving skills in particular. As the majority of online dictionaries are all freely accessible, learning how to use them efficiently will reduce inequalities among students coming from different socio-economic backgrounds.

After the completion of the project, we will be able to:

- give guidance on how to integrate (the teaching of) dictionary use into traditional classroom teaching in higher education, thus making the teaching and learning process more effective,
- provide background information on dictionaries,
- increase students' potential in achieving the goal of becoming autonomous learners,
- improve our students' level of English language proficiency,
- contribute to more students being able to successfully cope with university requirements,
- improve the digital competencies of students,
- contribute to the reduction of inequalities among students coming from different socio-economic backgrounds.

References

- Baron, N. S. (2021). *How we read now: strategic choices for print, screen, and audio*. Oxford University Press.
- Campoy-Cubillo, M. C. (2015). Assessing Dictionary Skills. *Lexicography ASIALEX*, 2, 119–141. <https://doi.org/10.1007/s40607-015-0019-2>
- Chan, A. Y. W. (2012). Cantonese ESL Learners' Use of Grammatical Information in a Monolingual Dictionary for Determining the Correct Use of a Target Word. *International Journal of Lexicography*, 25(1), 68–94. <https://doi.org/10.1093/ijl/ecr014>
- Gavriilidou, Z., & Konstantinidou, E. (2022). The effect of an explicit and integrated dictionary awareness intervention program on dictionary use strategies. In A. Klosa-Kückelhaus, S. Engelberg, Ch. Möhrs, & P. Storjohann (Eds.), *Dictionaries and Society: Proceedings of the XX EURALEX International Congress* (pp. 471–479). IDS-Verlag.

- Gonda, Zs. (2009). A szótárhasználati kompetencia elsajátítása és fejlesztése. [Acquiring and developing competence in dictionary use.] *Anyanyelv-pedagógia*. [Mother-tongue pedagogy], 2(2). <https://www.anyanyelv-pedagogia.hu/cikkek.php?id=160>
- Gouws, R. (2022). Dictionaries: Bridges, Dykes, Sluice Gates. In A. Klosa-Kückelhaus., S. Engelberg, Ch. Möhrs, & P. Storjohann (Eds.), *Dictionaries and Society: Proceedings of the XX EURALEX International Congress* (pp. 36–52). IDS-Verlag.
- Lew, R. (2010). Users take shortcuts: navigating dictionary entries. In A. Dykstra, & T. Schoonheim (Eds.), *Proceedings of the 14th EURALEX International Congress* (pp. 1121–1132). Fryske Akademy.
- Lew, R. (2013). Online Dictionary Skills. In I. Kosem, J. Kallas, P. Gantar, S. Krek, M. Langemets, & M. Tuulik (Eds.), *Electronic Lexicography in the 21st Century: Thinking Outside the Paper. Proceedings of the eLex 2013 Conference* (pp. 16–31). Institute for Applied Slovene Studies/Eesti Keele Instituut.
- Lew, R. (2016). Dictionaries for Learners of English. *Language Teaching*, 49(2), 291–294. <https://doi.org/10.1017/S026144481500049X>
- Margalitadze, T. (2018). Once Again Why Lexicography is Science. *Lexikos*, 28(1), 245–261. <https://doi.org/10.5788/28-1-1464>
- Margalitadze, T. (2020). Language and Ecology of Culture. *Lexicographica*, 36, 225–240. <https://doi.org/10.1515/lex-2020-0012>
- Margalitadze, T., & Meladze, G. (2023). How to Solve Problems in Dictionary Use: The Georgian Case. *Lexikos*, 33(2), 63–78. <https://doi.org/10.5788/33-2-1840>
- Márkus, K., & Szöllősy, É. (2006). Angolul tanuló középiskolásaink szótárhasználati szokásairól. [The Dictionary Use Habits of Hungarian Secondary School Students Learning English as a Foreign Language]. In T. Magay (Ed.), *Szótárak és használóik*. [Dictionaries and their Users.] (pp. 95–116). Akadémiai Kiadó.
- Nied Curcio, M. (2022). Dictionaries, Foreign Language Learners and Teachers. New Challenges in the Digital Era. In A. Klosa-Kückelhaus., S. Engelberg, Ch. Möhrs, & P. Storjohann (Eds.), *Dictionaries and Society: Proceedings of the XX EURALEX International Congress* (pp. 71–84). IDS-Verlag.
- P. Márkus, K. (2023). *Teaching Dictionary Skills*. L'Harmattan Kiadó.
- P. Márkus, K., Fajt, B., & Dringó-Horváth, I. (2023). Dictionary Skills in Teaching English and German as a Foreign Language in Hungary: A Questionnaire Study. *International Journal of Lexicography*, 36(2), 173–194.
- P. Márkus, K., & Dringó-Horváth, I. (2023). Developing Dictionary Skills through Monolingual and Bilingual English Dictionaries at Tertiary-level Education in Hungary. *Lexikos*, 33(1), 350–381. <https://doi.org/10.5788/33-1-1821>
- Roby, W. (2006). The Internet, Autonomy, and Lexicography: A Convergence? *Mélanges CRAPEL*, 28, 47–66.

Rundell, M. (1999). Dictionary Use in production. *International Journal of Lexicography*, 12(1), 35–53.

Rundell, M. (2016). Dictionaries and crowdsourcing, wikis and user-generated content. In P. Hanks, & G.-M. de Schryver (Eds.), *International Handbook of Modern Lexis and Lexicography* (pp. 1–16). Springer-Verlag.

Tarp, S., & Gouws, R. H. (2020). Reference Skills or Human-Centered Design: Towards a New Lexicographical Culture. *Lexikos*, 30, 470–498.

Vrbinc, M., & Vrbinc, A. (2006). A research-based study of foreign students' use of grammatical codes in five leading British learners' dictionaries. *Linguistica*, 46(1), 227–242.

Contact information

Tinatin Margalitadze

Ilia State University

tinatin.margalitadze@iliauni.edu.ge

Katalin P. Márkus

Károli Gáspár University of the Reformed Church in Hungary

p.markus.kata@kre.hu

Barbora Štěpánková, Lucie Poláková, Jana Šindlerová, and
Michal Novák

WHAT CAN DICTIONARIES TELL US ABOUT PRAGMATIC MARKERS

Building the Lexicon of Epistemic and Evidential Markers in Czech

Abstract In this paper, we explore the possibilities and challenges of lexicographic treatment of pragmatic markers, specifically epistemic and evidential markers in Czech. Our starting point is a detailed comparison of how these expressions are treated in contemporary monolingual Czech dictionaries. Following this, we present the development of the SEEMLex lexicon of Czech epistemic and evidential markers which is based on detailed annotation of selected expressions using data from a Czech-English parallel corpus. We describe the features we annotate when analysing the expressions studied, outline the main aspects that constitute or distinguish their meanings, and emphasise the importance of considering the communicative function in which these expressions are used. Additionally, we highlight the benefits of using a specialised lexical database for the lexicographic processing of pragmatic expressions in general. We demonstrate our approach with a draft of a dictionary entry for the common Czech epistemic marker *asi* ‘probably’ providing a comprehensive example of our methodology.

Keywords specialized dictionary; epistemic markers; communicative functions; annotation; Czech

1. Introduction

Compiling a monolingual dictionary is a complex task involving many conceptual decisions in order to ensure a comprehensive, yet comprehensible and consistent treatment of word meanings. In the Czech lexicographic tradition, the process of compiling a monolingual dictionary is usually carried out in alphabetical order¹, which in the long run can lead to difficulties in maintaining consistency of treatment due to the change of authors and subsequent changes in the principles of compilation. As a result, expressions belonging to the same semantic group may be treated differently.

This is particularly the case for expressions with a weakened lexical meaning, i.e., grammatical or functional words, and thus also for pragmatic markers, including epistemic and evidential markers (hereafter EEM). There are several

¹ This approach is described, e.g., in the unpublished guidelines for the compilation of the SSJČ dictionary (*Směrnice pro vypracovávání rukopisu Slovníku spisovného jazyka českého*, 1957). There appeared one notable exception to this approach – the comprehensive proposal for the delimitation and treatment of secondary prepositions, including their list and description of their meanings, which was comprised for and applied in the one-volume SSC dictionary.

reasons for this: 1) monolingual dictionaries focus primarily on content words and leave other expressions on the periphery of interest, or 2) it is only recently that the interest in pragmatic expressions and pragmatic meaning has arisen, thus, there is no standardised lexicographic treatment agreed upon yet, or 3) Czech pragmatic expressions are often homonymous/polysemous and in many cases it is difficult to adequately distinguish the epistemic meanings from other meanings.

The aim of this study is to present a proposal for an architecture of a lexicon of Czech epistemic and evidential markers that would provide a unified and complex description of their meaning.² First, we define the group of Czech EEMs and comment on the existing approaches to their description, including lexicographic approaches. Then, the process of building the SEEMLex lexicon based on the underlying annotation of selected EEMs in corpus data is presented. As a case study, we provide the results of the annotation of the most frequent Czech epistemic marker *asi* ‘probably’, and comment on the relevant semantic and pragmatic features to be captured in its lexicographic description. Also, we present a sample lexicon entry. Finally, we summarise the advantages of a specialised lexical database as a basis for the treatment of a specific class of words in a monolingual dictionary.

2. Epistemic and Evidential Markers

Pragmatic markers, in general, are expressions that encode people’s opinions, assumptions and beliefs regarding the propositional content. Their meaning is primarily semantico-pragmatic in nature. In the Czech linguistic tradition, they are classified as *particles*.³ EEMs are then traditionally referred to as *epistemic particles* (e.g., Komárek et al., 1986; Cvrček et al., 2010).

In the literature, the relationship between epistemic and evidential markers is perceived diversely (cf., e.g., de Haan, 2001; Nuyts, 2001; Plungian, 2001). Our approach assumes a general overlap of both categories in accordance with some theoretical approaches (cf. Hoyer, 2008; Carretero & Zamorano-Mansilla, 2013; Komárek et al., 1986; Cvrček et al., 2010) and with the support of our pilot analysis (cf. Šindlerová et al., 2023). Thus, in the context of a project focused on exploring EEMs, we understand them as one broader group containing permeable categories of epistemic markers (such as *možná* ‘maybe’,⁴ *určitě* ‘certainly’), evidential markers (*údajně* ‘allegedly’, *podle všeho* ‘to all appearances’) and markers confirming/emphasizing the speaker’s strong belief in their being right, the so-called confirmatory expressions (cf. Rozumko, 2016) (*ovšem* ‘of course’).

² For EEMs, we use the term *meaning* in the sense of the term *relational meaning* (cf. Filipec & Čermák, 1985, p. 39).

³ In Czech linguistics, *particles* are defined as expressions with a predominant attitudinal function; they are generally accepted as one of ten parts of speech. For a detailed analysis of the concept of *particles* in Slavic languages compared to anglophone linguistics using the example of epistemic adverbs, see Rozumko (2016). For a description of the behaviour of these expressions, cf. also Volkova (2017), or Grochowski et al. (2014).

⁴ Here are the most common translation equivalents. For specific examples, equivalents appropriate to the context are used.

2.1 Dictionary Treatment of the EEMs

As mentioned above, in Czech monolingual dictionaries, EEMs are usually treated as function words. As such, they have received limited attention so far. The definitions of EEMs mostly comprise only a vague description of the degree of certainty (1).

- (1) *dozajista*: ‘vyjadřuje nejvyšší míru jistoty, přesvědčení mluvčího o platnosti tvrzení’ (*certainly*: ‘expresses the maximum degree of certainty, speaker’s belief in the validity of the proposition’) (ASSČ)

Only rarely and inconsistently do we come across attempts to capture other aspects (2). In some grammars, their role in expressing negativity (e.g., Daneš et al., 1987) or their function in communication (Hoffmannová et al., 2019) is mentioned.

- (2) *zajisté*: ‘2. vyjadřuje zdůrazněně souhlas; 3. vyjadřuje subjektivní přesvědčení o něčem’ (*certainly*: ‘2. emphatically expresses agreement, 3. expresses a subjective belief about something’) (SSJČ)⁵

This does not correspond to the definitions in monolingual dictionaries of languages other than Czech which often provide more detailed explanations (cf. selected definitions of *maybe* (3)), commenting specifically on the possible intentions of the speaker or giving hints on the degree of politeness.

- (3) *maybe*: ‘used to politely suggest or ask for something; used to avoid giving a clear or certain answer to a question’ (CLD)

Therefore, in general, Czech lexicographic treatments of EEMs typically contain only a basic semantic feature, namely the aforementioned specification of the degree of certainty, which most closely resembles the treatment of content words. However, as “words with a primarily pragmatic nature and usage”⁶ (Čermák, 1992, p. 257), the expressions under examination require a more specific approach, based on the pragmatic component of their meaning. Our pilot study (Štěpánková et al., 2023) suggested that the degree of certainty may be weakened in some uses and that the **communicative function** (CF) of an utterance can serve as a crucial component to follow. According to Grepl (2017), we understand CF as “the meaning of an utterance resulting from the intention with which the utterance is produced by the speaker towards the addressee in that particular communicative situation.”⁷ Based on our experience with empirical data so far, we hypothesise that an annotation of CF is crucial not only for the description of basic EEM meanings but also for distinguishing other meanings of the studied expressions. A large-scale CF annotation can provide evidence for the repeated and regular use of some of these meanings and thus for their lexicalization.

⁵ The SSJČ and SSČ dictionaries are cited as <https://prirucka.ujc.cas.cz/>

⁶ “...slova s primárně pragmatickou povahou a územ...” (Čermák, 1992, p. 257)

⁷ “...smysl výpovědi vyplývající ze záměru, s jakým je nějaká výpověď mluvčím vůči adresátovi v dané konkrétní komunikační situaci produkována...” (Grepl, 2017)

3. Methodology

The specialised SEEMLex lexicon is planned as one of the outputs of the project researching EEMs. In recent years, electronic lexicons describing functionally defined groups of expressions have often been created in diverse areas, e.g., in the area of subjectivity (Veselovská & Bojar, 2013), or valency (Lopatková et al., 2016); in the area of discourse markers and connectives, see e.g., Mírovský et al. (2017) for Czech, or Stede (2002) for German; a multilingual database of discourse markers is available online (Stede et al., 2019)⁸; in the area of epistemic and evidential markers, see e.g., proposal for a database by Wiemer & Stathi (2010). Such lexicons are usually created on the basis of large corpus data and their specific annotation, which is also the case of the SEEMLex.

In our project, we use the parallel InterCorp v15 corpus (Čermák & Rosen, 2012), specifically the core part of its Czech and English sections, containing mainly fiction, as the underlying language resources for lexicon development. The fiction data were chosen as suitable for the study of EEMs for the following reasons:

- presumed closeness to spoken language (e.g., in terms of high frequencies of epistemic markers, a certain degree of subjectivity);
- in contrast to originally spoken texts, the fiction writing also mostly offers elaborate situational context which is helpful for EEM interpretation;⁹
- unlike often in journalistic texts, the origin of a fiction text (the author and the original language) is known. Our primary focus is on original Czech texts, and, secondly, original English texts translated into Czech;
- reliable translation quality of the texts is another advantage of using a parallel corpus. It allows us to use translation equivalents to clarify the meaning of the Czech markers in context (cf. Aijmer et al., 2006).

In contrast to several studies devoted to epistemic modality (cf. the Modal Corpus described in Pietrandrea (2018)), SEEMLex does not use a corpus-driven approach to identify EEMs, instead, we annotate a predetermined list of them.

This headword list has been compiled by a manual selection of lemmas and forms from various Czech grammars, complemented by a selective list of markers annotated as modal or attitudinal in the Prague Dependency Treebanks. While grammars mostly list typical (very frequent) expressions representing the epistemic group and usually do not provide any context, the corpora used to compile the headword list (PDT 3.5, Hajič et al., 2018, and PDTSC, Mikulová et al., 2017) capture various expressions in their (syntactic) contexts, and additionally enrich the list with less frequent items. The comprehensive list contains approximately 140 entries. It includes both single-word (*pravděpodobně* ‘probably’) and multi-word (*s jistotou* ‘certainly’, lit. ‘with certainty’) markers. We do

⁸ <http://connective-lex.info/>

⁹ On the other hand, the written texts have the disadvantage of not including intonation, which often serves as a very useful interpretative device in the case of particles.

not include modal verbs – their properties are somewhat different, and, in contrast to EEMs, they are well researched in Czech (cf. Grepl, 1979; Ivanová, 2017, etc.).

3.1 Tool

Given the headword list and the parallel corpus, we collect all occurrences of the expressions in the corpus and select samples of them for manual annotation. The manual annotation of the EEMs is conducted using the TEITOK web-based platform (Janssen, 2016). For each sampled expression, the annotators can see the sentence it appears in, including the possibility to explore an arbitrarily large context of the sentence, which may be crucial for the annotation of modality.

Furthermore, the annotation environment can display the English equivalent of the sentence and highlight the expression's counterpart in the sentence. While the sentence alignment is an integral part of the InterCorp corpus, the counterpart of the expression is obtained automatically by running the AWESOME aligner (Dou & Neubig, 2021). We fine-tuned the default model based on multilingual BERT (Devlin et al., 2019) first on the parallel Czech-English data from PCEDT 2.0 (Hajič et al., 2012) in the unsupervised setting, followed by Czech-English manual word alignments (Mareček, 2008) in the supervised setting. While annotating the modality features, the annotators are also asked to fix potential errors in the automatic alignment of the expressions.

The annotators may also decide to label an annotated sentence as a candidate for dictionary exemplification. Out of these candidate sentences, we manually select the most suitable examples to be shown in the SEEMLex lexicon.

3.2 Annotation of EEM Features

Each instance of the selected markers in the corpus data is annotated for a set of features. This repertoire of features was compiled based on the available state-of-the-art studies (e.g., Wiemer & Stathi, 2010; Lavid et al., 2016; Pietrandrea, 2018) and was confirmed convenient through test annotation (Štěpánková et al., 2023).

Although our project focuses on expressions with epistemic and evidential meanings, we consider it necessary to annotate also other meanings of the given expressions at least in a basic manner. In this way, we are able to document their widely polysemous/homonymous nature and map their overall use. For example, for the expression *jistě* 'certainly', the meaning of a manner adverb – 'walk surely' (4) or response particle 'sure' (5) are annotated.

(4) *Kráčela lehce a **jistě**.*¹⁰ 'She walked lightly and **surely**.'

(5) "Chodil jste do kostela?" "**Jistě**. Každé Vánoce a Velikonoce." "Did you go to church?"
"Sure. Every Christmas and Easter."

¹⁰ All examples used come from the InterCorp corpus.

In agreement with Carretero & Zamorano-Mansilla (2019), we also consider as EEMs those uses where an additional feature is present in combination with the epistemic meaning, e.g., expressing an attitude (6).

- (6) “Znáte **jistě** tuto scénu z desítek špatných filmů: hoch a dívka se drží za ruce a běží krásnou jarní (eventuelně letní) přírodou. ‘You **certainly** remember this scene from dozens of bad films: a boy and a girl are running hand in hand in a beautiful spring (or summer) landscape.’

The annotated features (see Table 1) relate to the expression itself (e.g., position in the sentence), describe phenomena in its close context (grammatical features of the predicate, presence of evidentiality, negation, contrast, etc.), or comment on the utterance as a whole (CF).

Table 1: List of annotated features

Annotated feature	Values
Type of use	epistemic, evidential, confirmatory, response, other, autosemantic
Degree of certainty	high, higher medium, medium, low
Type of CF	assertive, directive/contact, interrogative, commissive, (dis)approval, expressive
Specific CF	e.g., assumption, recommendation, wish
Scope	clause/member
Predicate verb	verb tag
Position in a sentence	first, last, other
Negation	Y/N
In a contrastive pattern	Y/N
Other modal expression	e.g., intensifier, modal marker, modal verb
Type of evidence	sensory, hearsay, reasoning, inference
Translation equivalent	choice from the parallel English sentence

4. Case Study – *asi*

Asi is the most frequent epistemic marker on our list, therefore we have selected it as an example lexicon entry. Moreover, it is one of the expressions that have already been processed in the most recent (unfinished) monolingual dictionary, the ASSČ¹¹, which itself is proclaimed to be based on corpus data and its concept explicitly mentions a shift towards reflecting pragmatics in the processing of entries (cf. Kochová & Opavská, 2016).

In this section, we first compare and critically evaluate three different ways of lexicographic treatment of the selected marker in Czech dictionaries, then we present our draft for the EEM lexicon entry, and, finally, we discuss the underlying principles in detail.

¹¹ Currently, entries starting with the first ten letters of the Czech alphabet (A–CH) have been published, i.e., approx. 20,000 entries.

The expression *asi*, roughly translatable as ‘probably’, is considered an epistemic marker situated approximately in the middle of the certainty scale (in agreement with e.g., Komárek et al., 1986; or Grepl, 2017). This basic semantic characteristic is also evident in dictionary treatments. In older monolingual dictionaries (SSJČ (7), and SSČ (8)), the word is described by means of two meanings, or shades of meaning, both expressing a lower degree of certainty. In the SSČ, the only meaning included applies the lower degree of certainty also to the meaning of approximation (paraphrased by *přibližně*).

- (7) SSJČ: **1.** *přibližně*: a. před týdnem; a. pět knih; a. čtyři lidé; **2.** *jak se zdá; pravděpodobně, snad, možná, patrně*: to a. nepůjde; a. to tak je; a. někde přišlo; a. to přinesu¹²
- (8) SSČ: *vyj. menší míru jistoty, pravděpodobně, možná, snad 1, patrně*: asi bude pršet; asi tak před týdnem *přibližně*¹³

In the most recent ASSČ dictionary (9), *asi* is divided into three separate meanings: Meaning 1 contains uses expressing at least a medium degree of certainty; Meaning 2 expresses the adverbial meaning of measure (see (7), Meaning 1 above), and Meaning 3 records other uses expressing various sentiments of the speaker towards the proposition. It must be noted that in the last type of meaning, the original certainty meaning is no longer manifested and the attitudinal function prevails. To sum up, in the ASSČ dictionary, we can see a noticeable attempt to separate the certainty meaning from other types of meaning.

- (9) ASSČ: **1.** *vyjadřuje střední nebo vyšší míru jistoty syn. pravděpodobně*: Soupeři se ho asi bojí. Sluneční hodiny zná asi každý. Krupobití a vichru se asi nevyhneme. *ve funkci citoslovce* Pojedete na dovolenou? – Asi. *asi ano*; **2.** *vyjadřuje přibližnost míry, množství, délky trvání, zprav. před číselným výrazem syn. přibližně*: Seskočil z výšky asi jednoho a půl metru. Na letišti čekalo asi pět set lidí. Pracovala tam asi rok. Zpoždění vlaku bude asi šedesát minut. **3.** *zdůrazňuje citový postoj mluvčího k situaci •(v otázce) zvědavost, nevědomost*: Kdepak mám asi lístek? Budeme mít novou paní učitelku. Jaká asi bude? *•expresivní rozhořčení, nesouhlas, často při odmítání předchozí výpovědi*: Ty ses asi zbláznil! Byl tady Tom! – Kde by se tady asi vzal? Jestli toho nenecháš, šeredně na to doplatíš. – A co mám asi dělat?¹⁴

¹² ‘**1.** *approximately*: a. a week ago; a. five books; a. four people; **2.** *it seems, probably, perhaps, maybe, apparently*: it seems impossible; it seems so; it has probably rained somewhere; I may bring it with me’)

¹³ ‘*expressing lower degree of certainty, probably, maybe, perhaps 1, apparently*: it may rain; approximately a week ago *approximately*’

¹⁴ ‘**1.** *expressing medium or higher degree of certainty, synonymous with probably*: His enemies probably fear him. Sundial is known by probably everyone. We probably won’t avoid hail and wind. *in the function of an interjection*: Are you going on vacation? – Probably. *Probably we are*. **2.** *expressing inexactness of the measure, quantity, duration, usually in front of a numerical expression, synonymous with about*: He jumped from a height of about one and a half meters. There were about 500 people waiting at the airport. She has been working there for about a year. The train will be delayed by about sixty minutes. **3.** *emphasizing the emotional attitude of the speaker to the situation •(in questions) curiosity, ignorance*: Where might my ticket be? We will have a new teacher. What will she be like? *•expressive indignation, disagreement, often when rejecting previous statement*: You must be crazy! Tom was here! – How would he have gotten here? If you don’t stop, you’ll pay dearly for it. – And what am I supposed to do?’

On the other hand, the authors' lack of experience or methodological lexicographic support in the processing of expressions with a strong pragmatic component of meaning is evident here, as well as insufficient comparative approach that would take into account also other expressions from this functional-semantic group.

Meaning 1 also includes examples that are defined by their communicative function rather than by the degree of certainty, e.g., the conversational premise *Sluneční hodiny zná asi každý*. 'Sundial is known by probably everyone.' or the response particle *Pojedete na dovolenou? – Asi*. 'Are you going on vacation? – Probably.'

What is more, Meaning 3 merges different types of attitudes which imply different synonymous alternatives, e.g., in the example *Ty ses asi zbláznil* 'You must be crazy' [lit. 'You have gone probably crazy'], *asi* can be replaced by a number of other certainty markers with varying degrees of certainty (*určitě* 'definitely', *nejspíš* 'probably'); on the other hand, *Kde by se tady asi vzal?* 'How would he have [lit. 'probably'] gotten here?' is a use in which no such substitution is possible. It is therefore worth considering whether these examples should be grouped together under a single meaning.

Within the case study, we performed a parallel annotation of 200 occurrences of *asi* (100 from original Czech texts, 100 from the Czech translations of English originals). Randomly selected samples were annotated in parallel by three annotators, native speakers of Czech with a linguistic background. Inter-annotator comparisons were made on the annotated data, focusing mainly on the basic type of use categories (epistemic – pragmatized – autosemantic). Inconsistencies – in most cases – included annotator's misinterpretation of the annotation guidelines, the treatment of expressive usages with a more vague solution in the guidelines, multiple possible interpretations of a sentence. The first two types of inconsistencies should be improved in future annotations, on the other hand, different interpretations are unavoidable. Our draft of an EEM lexicon entry is based on the analysis of the annotation results as well as on the dictionary comparison above. When creating a dictionary entry, we are primarily guided by two principles: 1. the presence or absence of epistemic modality or its weakening, and 2. the communicative function in which the expression is used.

In our draft, the basic, unmarked use of *asi* is the epistemic meaning with a medium certainty or higher medium certainty degree. Within this major use, several communicative functions can be distinguished. The strongest one is the assertive CF which includes several subtypes of CFs mostly distinguishable thanks to the lexical or syntactic contexts. The default assertive CF is the assumption (I.a). Further, if the context contains strong evidence, the CF of the utterance is explanation (I.b); different verb tenses and moods imply the CFs of future guess (I.c). Within the epistemic type of use, we have also documented a directive/contact CF subtype which is primarily used to express a recommendation (II.a), but is also further used as a conversational formula/assumption (II.b). Meaning III shows a specific type of assertive CF with a weakened degree of certainty – introspection. In Meanings IV–VII, *asi* appears in functions of attitudinal pragmatic markers. Meaning VIII describes various uses of approximation.

Draft of a SEEMLex entry *asi*¹⁵

Epistemický význam

I. Vyjádření střední či vyšší střední jistoty vzhledem k propozici

I.a Domněnka o tom, co se stalo nebo děje: Šlo **asi** o import z Dálného východu.

I.b Vysvětlení, s přítomností evidence: **Asi** jsem na chvíli usnul, protože když jsem otevřel oči, byl jsem ve třídě s provaleným stropem sám.

I.c Odhad toho, co se bude dít: Ti vojáci tu **asi** taky nebudou věčně.

II. Direktivní/kontaktoý význam, epistemičnost částečně oslabená

II.a Doporučení: „**Asi** byste měl,“ řekl.

II.b Direktivní předpoklad, předjímání názoru komunikačního partnera: **Asi** se tomu divíš, takhle přímo jsi to ode mne neslyšel.

III. Introspekce, epistemičnost částečně oslabená: Já se v té chvíli **asi** pomát.; **Asi** bych si od něho měla něco přečíst.

Bez epistemičnosti, pragmatikalizovaný význam

IV. Tázací: Jaké poruchy by se **asi** jevily u profesora Devrienta, kdybych mu odňal pravý čelní mozkový lalok?

V. (Slabý) souhlas/nesouhlas: „To je všechno, co máme? Fazole?“ „**Asi**.“

VI. Hedging, konverzační formule: „Je mi hrozně líto, ale to **asi** nepůjde.“

VII. Expresivní, umocňuje postoj: To bude to středisko **asi** pěkně vypadat.

Bez epistemičnosti, plnovýznamové

VIII. Přibližnostní význam

VIII.a Přibližné množství, před číselným výrazem nebo srovnáním: večer **asi** v pět nebo v šest; Byl velký **asi** jako menší město.

VIII.b Přibližnost, podobnost: Ale říkal **asi** tohle.

‘Epistemic modality

I. expression of medium and higher medium certainty regarding a proposition

I.a Assumption about past or present events: This **probably** referred to silk imported from the Far East.

I.b Explanation: I **must** have nodded off, because when I opened my eyes I was alone in the classroom with the collapsed ceiling.

I.c Estimation of what will happen: Those soldiers **might** not be here forever either.

II. Directive/contact meaning, epistemic meaning partially weakened

¹⁵ The SEEMLex lexicon will be provided both in Czech and English language version.

II.a Recommendation: “*Perhaps* you should,” he said.

II.b Directive assumption, anticipating the partner’s opinion: *Maybe* that comes to you as a bit of a shock, you’ve never heard it like that, straight from me.

III. Introspection, epistemic meaning partially weakened: *And at that very instant I must* have gone mad.; *Maybe* I should read one of his things.

Without epistemicity, pragmaticalized meaning, politeness, hedging

IV. Interrogative: *What kind of disturbances would appear in Professor Devrient if I removed his right frontal lobe?*¹⁶

V. (Weak) agreement/disagreement: “*Is that all we have? Beans?*” “*Could be.*”

VI. Hedging, some form of weakening, conversational figure: ‘*Oh, I’m terribly sorry, but I don’t think that’s possible.*

VII. Expressive, emphasizes an attitude: *What an awful place that re-education centre must be!*

Without epistemicity, autosemantic

VIII. Approximation

VIII.a Approximate quantity preceding a numerical expression or comparison: *in the evenings, around* five or six; *It was the size of a small city.*

VIII.b Approximation, similarity: *But what he said was roughly* this.’

5. Conclusion

By comparing traditional lexicographic approaches to pragmatic markers, we have demonstrated that contemporary Czech dictionaries typically handle them using principles more suited to content words, i.e., the pragmatic component is often disregarded. The example of the relatively detailed treatment of *asi* in the ASSČ dictionary demonstrates the current lack of theoretical lexicographic support for dealing with such expressions.

Hence, the primary objective of this study was to propose a more comprehensive lexicographic approach to EEMs in a specialized lexicon based on an in-depth annotation of the studied expressions in context using a parallel corpus. In the annotation process, two guiding principles for the lexicographic treatment of EEMs proved relevant: the first one is the degree of certainty conveyed by a given expression in a specific context, the second one is the communicative function of the utterance containing the expression. While the degree of certainty (epistemicity) may be weakened or completely emptied for some uses, the communicative function remains a strong lead to distinguish among various attitudinal meanings of the examined expressions.

¹⁶ In the Meanings IV, VI, and VIII.a, *asi* is implied or paraphrased by quite different linguistic features in the English InterCorp translation (e.g., in VI lit. *perhaps not possible* = *I don’t think that’s possible*).

The future SEEMLex lexicon will offer a comprehensive overview of the selected items. Each entry will primarily focus on describing the epistemic and evidential features of a given marker but it will also capture other functions (e.g., response functions, expressing politeness, etc.), as well as intrinsic lexical meanings (e.g., expressing approximation, manner). Apart from ensuring greater consistency, this approach facilitates an in-depth analysis of the universal and specific features of the individual expressions within the given group including their mutual relations (such as synonymy, antonymy, etc.).

References

- Aijmer, K., Foolen, A., & Simon-Vandenberg, A-M. (2006). Pragmatic markers in translation: a methodological proposal. In K. Fischer (Ed.), *Approaches to discourse particles*. (pp. 101–114). Elsevier.
- ASSČ: *Akademický slovník současné češtiny*. Retrieved May 26, 2024, from <https://slovníkcestiny.cz/>
- CLD: *Cambridge Learner's Dictionary*. Retrieved January 24, 2024, from <https://dictionary.cambridge.org/dictionary/english/maybe>
- Carretero, M., & Zamorano-Mansilla, J. R. (2013). Annotating English adverbials for the categories of epistemic modality and evidentiality. In J. Marín Arrese, M. Carretero, J. Arús, & J. van der Auwera (Eds.), *English Modality: Core, Periphery and Evidentiality. Topics in English Linguistics*, 81. (pp. 317–355). Mouton De Gruyter.
- Carretero, M., & Zamorano-Mansilla, J. R. (2019). Disentangling epistemic modality, neighbouring categories and pragmatic uses: the case of English epistemic modal adverbs. In K. Filippi-Deswelle (Ed.), *Quinze études de cas sur les modalités linguistiques. / Fifteen case studies on types of linguistic modalities* (pp. 131–157). Publications Électroniques de l'ERLAC, Epilogos 6.
- Cvrček, V., Kodýtek, V., Kopřivová, M., Kovářiková, D., Sgall, P., Šulc, M., Táborský, J., Volín, J., & Waclawicová, M. (2010). *Mluvnice současné češtiny I*. Karolinum.
- Čermák, F. (1992). Paradigmatika a syntagmatika slovníku: problémy a možnosti. *Slovo a slovesnost*, 53(4), 249–264.
- Čermák, F., & Rosen, A. (2012). The case of InterCorp, a multilingual parallel corpus. *International Journal of Corpus Linguistics*, 17(3), 411–427.
- Daneš, F., Hlavsa, Z., & Grepl, M. (Eds.). (1987). *Mluvnice češtiny 3*. Academia.
- De Haan, F. (2001). The Relation between Modality and Evidentiality. *Linguistische Berichte*, 9, 201–216.
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In J. Burstein, Ch. Doran, & Th. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the*

Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 4171–4186). Association for Computational Linguistics.

Dou, Z., & Neubig, G. (2021). Word Alignment by Fine-tuning Embeddings on Parallel Corpora. In P. Merlo, J. Tiedemann, & R. Tsarfaty (Eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 2112–2128). Association for Computational Linguistics.

Filipec, J., & Čermák, F. (1985). *Česká lexikologie*. Academia.

Grepl, M. (2017). Komunikační funkce výpovědi. In P. Karlík, M. Nekula, & J. Pleskalová (Eds.), *CzechEncy – Nový encyklopedický slovník češtiny*. Retrieved May 24, 2024, from <https://www.czechency.org/slovník/KOMUNIKA%C4%8CN%C3%8D%20FUNKCE%20V%C3%9DPOV%C4%9ADI>

Grepl, M. (1979). Úvodní poznámky k tzv. jistotní modalitě. *Slovo a slovesnost*, 40(2), 81–87.

Grochowski, M., Kisiel, A., & Żabowska, M. (2014). *Słownik gniazdowy partykuł polskich*. Polska Akademia Umiejętności.

Hajič, J., Hajičová, E., Panevová, J., Sgall, P., Bojar, O., Cinková, S., Fučíková, E., Mikulová, M., Pajas, P., Popelka, J., Semecký, J., Šindlerová, J., Štěpánek, J., Toman, J., Urešová, Z., & Žabokrtský, Z. (2012). Announcing Prague Czech-English Dependency Treebank 2.0. In N. Calzolari, K. Choukri, Th. Declerck, Mehmet U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)* (pp. 3153–3160). European Language Resources Association (ELRA).

Hajič, J., Bejček, E., Bémová, A., Buráňová, E., Hajičová, E., Havelka, J., Homola, P., Kárník, J., Kettnerová, V., Klyueva, N., Kolářová, V., Kučová, L., Lopatková, M., Mikulová, M., Mirovský, M., Nedoluzhko, A., Pajas, P., Panevová, J., Poláková, L., Rysová, M., Sgall, P., Spoustová, J., Straňák, P., Ševčíková, M., Štěpánek, J., Urešová, Z., Vidová Hladká, B., Zeman, D., Zikánová, Š., & Žabokrtský, Z. (2018). *Prague Dependency Treebank 3.5*. ÚFAL, LINDAT/CLARIN, Charles University. <http://hdl.handle.net/11234/1-2621>

Helcl, M., Sochová, Z., & Kozlová, K. (Eds.). (1957). *Směrnice pro vypracovávání rukopisu Slovníku spisovného jazyka českého*. Interní tisk. ÚJČ.

Hoffmannová, J., Homoláč, J., & Mrázková, K. (Eds.). (2019). *Syntax mluvené češtiny*. Academia.

Hoye, L. F. (2008). Evidentiality in discourse: A pragmatic and empirical account. In J. Romero-Trillo (Ed.), *Pragmatics and Corpus Linguistics* (pp. 151–174). De Gruyter Mouton.

Ivanová, M. (2017). *Modálnost a modálne verbá ve slovenčině*. FF Prešovskej univerzity.

Janssen, M. (2016). TEITOK: Text-Faithful Annotated Corpora. In N. Calzolari, K. Choukri, Th. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (pp. 4037–4043). ELRA.

- Kochová, P., & Opavská, Z. (Eds.). (2016). *Kapitoly z koncepce Akademického slovníku současné češtiny*. Ústav pro jazyk český AV ČR.
- Komárek, M., Kořenský, J., Petr, J., & Veselková, J. (Eds.). (1986). *Mluvnice češtiny 2*. Academia.
- Lavid, J., Carretero, M., & Zamorano-Mansilla, J. R. (2016). A linguistically-motivated annotation model of modality in English and Spanish: Insights from MULTINOT. *Linguistic Issues in Language Technology*, 14, 1–35. <https://aclanthology.org/2016.lilt-14.4/>
- Lopatková, M., Kettnerová, V., Bejček, E., Vernerová, A., & Žabokrtský, Z. (2016). *Valenční slovník českých sloves VALLEX*. Karolinum.
- Mareček, D. (2008). *Automatic Alignment of Tectogrammatical Trees from Czech-English Parallel Corpus*. [Master thesis. Charles University, MFF].
- Mikulová, M., Bémová, A., Hajič, J., Hajičová, E., Ircing, P., Kolářová, V., Lopatková, M., Mareček, D., Mírovský, J., Nedoluzhko, A., Pajas, P., Panevová, J., Peterek, N., Romportl, J., Sgall, P., Ševčíková, M., Štěpánek, J., Urešová, Z., & Žabokrtský, Z. (2017). *Prague Dependency Treebank of Spoken Czech 2.0*. ÚFAL, LINDAT/CLARIN, Charles University. <http://hdl.handle.net/11234/1-3189>.
- Mírovský, J., Synková, P., Rysová, K., & Poláková, L. (2017). CzeDLex – A Lexicon of Czech Discourse Connectives. *The Prague Bulletin of Mathematical Linguistics*, 109, 61–91.
- Nuyts, J. (2001). *Epistemic Modality, Language, and Conceptualization: A cognitive-pragmatic perspective*, Vol. 5. John Benjamins Publishing.
- Pietrandrea, P. (2018). Epistemic constructions at work. A corpus study on spoken Italian dialogues. *Journal of Pragmatics*, 128(15), 171–191.
- Plungian, V. (2001). The place of evidentiality within the universal grammatical space. *Journal of Pragmatics*, 33, 349–357.
- Rozumko, A. (2016). Linguistic Concepts across Languages: The Category of Epistemic Adverbs in English and Polish. *Yearbook of the Poznan Linguistic Meeting*, 2(1), 195–214.
- Stede, M. (2002). DiMLex: A Lexical Approach to Discourse Markers. In A. Lenci & V. Di Tomaso (Eds.), *Exploring the Lexicon - Theory and Computation*. Edizioni dell'Orso.
- Stede, M., Scheffler, F., & Mendes, A. (2019). Connective-lex: A web-based multilingual lexical resource for connectives. *Discours*, 24. DOI: 10.4000/discours.10098
- SSČ: Filipec, J., Daneš, F., Machač, J., Kroupová, L., Mejstřík, V., Poštolková, B., & Sochová, Z. (Eds.). (2003). *Slovník spisovné češtiny pro školu a veřejnost*. Academia.
- SSJČ: Havránek, B., Bělič, J., Helcl, M., Krístek, V., & Trávníček, F. (Eds.). (1960–1971). *Slovník spisovného jazyka českého*. ČSAV.
- Šindlerová, J., Štěpánková, B., & Andrén, I. L. (2023). Epistemická částice zřejmě pohledem paralelního korpusu. *Korpus – gramatika – axiologie*, 27, 37–52.

Štěpánková, B., Šindlerová, J., & Poláková, L. (2023). The Epistemic Marker *určitě* in the Light of Corpus Data. *Jazykovedný časopis*, 74, 130–139.

Veselovská, K., & Bojar, O. (2013). *Czech SubLex 1.0*. LINDAT/CLARIAH-CZ, ÚFAL.

Volkova, L. (2017). Pragmatic markers in dialogical discourse. *Lege artis. Language Yesterday, Today, Tomorrow*, 2(1), 379–427.

Wiemer, B., & Stathi, K. (2010). The database of evidential markers in European languages. A bird's eye view of the conception of the database. *Language Typology and Universals*, 63, 275–289.

Acknowledgements

The work has been supported by Czech Science Foundation under the project GA23-05240S and by Charles University Research Centre program No. 24/SSH/009. We gratefully thank Maarten Janssen, the main developer of TEITOK, for adapting the tool to meet the needs of this research.

Contact information

Barbora Štěpánková

Charles University, Prague
stepankova@ufal.mff.cuni.cz

Lucie Poláková

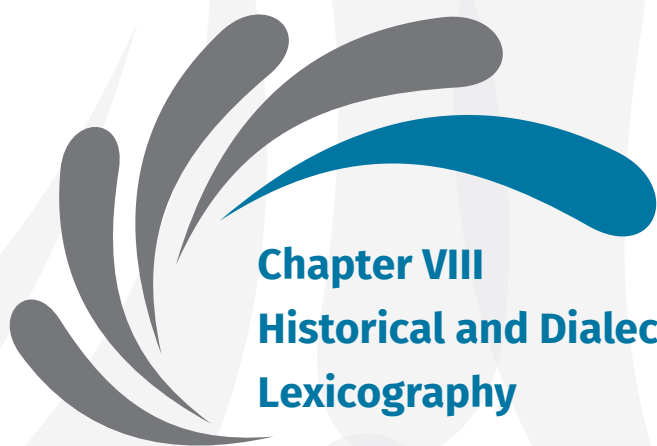
Charles University, Prague
polakova@ufal.mff.cuni.cz

Jana Šindlerová

Charles University, Prague
jana.sindlerova@ff.cuni.cz

Michal Novák

Charles University, Prague
mnovak@ufal.mff.cuni.cz



Chapter VIII
Historical and Dialect
Lexicography

David Lindemann and Mikel Alonso

LINKING HISTORICAL CORPUS DATA AND ANNOTATIONS USING WIKIBASE

Abstract This software demonstration presents a data model and a first use case for the representation of text corpus data on a Wikibase instance, including morphosyntactic, semantic and philological annotations as well as links to dictionary entries. Wikibase, an extension of MediaWiki, is the software that underlies Wikidata (Vrandečić & Krötzsch, 2014), an exceptionally large crowdsourced queryable knowledge graph, which includes nodes for ontological concepts, on the one hand, and for lexemes, lexeme senses and lexeme forms, on the other, together with annotations to and relations between them. We argue that the proposed model and the chosen software solutions for the representation of corpus and dictionary data, all free and open source, meet with the requirements of provenance transparency, open access and re-use, and the capability of collaborative work on the data. We also present our own scripts wrapped in a web application that shortcut several workflow steps in a first use case, a 1737 Basque manuscript, transcribed on Wikisource, and represented as an annotated dataset on our Wikibase instance.

Keywords Basque; historical corpus; Wikibase; corpus annotations; Linked Data

1. Introduction

With regard to the digitisation of historical texts in Basque, we have seen many efforts in recent years; various initiatives using different objectives and methodologies. Among these projects, we can find *Klasikoen Gordailua*¹, a repository of classic Basque literature. Another project, *Corpus Historikoa*², is a POS-lemma annotated corpus partly based on the former, enhanced with an online interface, allowing lemma-based searches. The corpus contains classical texts from 1545, when the first book written entirely in Basque was published, to 2000. The most recent project is the *BIM-SAHCOBA* corpus (Estarrona et al., 2022)³, a morphosyntactically annotated historical corpus which comprises the most significant works from the 15th century to the mid-20th century, when present-day standard Basque appeared. An interface allows searches based on the part of speech, on inflection form descriptions, lemmata, and neighbouring tokens and their metadata. An initiative with a totally different scope is the digital edition of the *Lazarraga Manuscript*⁴ (Bilbao et al., 2011), a recently found 16th century text, published along with philological annotations. It is our aim to propose a model which would combine contents of the cited versions of the Basque Historical Corpus, in a way that the user could also follow each corpus token back to its original source, e.g., a manuscript or early modern print facsimile image collection.

¹ Accessible at <https://klasikoak.armiarma.eus/>

² Accessible at <https://www.ehu.eus/etc/ch/>

³ Accessible at <http://bim.ix.eus/>

⁴ Accessible at <https://www.ehu.eus/monumenta/lazarraga/>

We have presented experiments to represent and interrelate historical and standard lexicographic data in Basque as Linked Data (Lindemann & Alonso, 2021), in which we have integrated a lexicographical dataset into the Wikisource and the Wikidata platforms. The resulting connections between a Basque historical source and the Wikidata knowledge graph, more precisely, connections between entries in a classic dictionary (Larramendi, 1745) on Wikisource and lexical entries (entities of type *lexeme*) describing standard Basque lemmata, can be manually or programmatically queried for, accessed and edited.

This software demonstration presents a data model and a first use case for the representation of contents of text corpus data on a Wikibase instance, including the types of annotations found in the above cited digital versions of historical Basque text, i.e., morphosyntactic, semantic and philological annotations, as well as links to dictionary entries.

2. Technical Considerations

Proposing the model described here, it is our goal to ensure interoperability with widely used standards for corpus annotation and lexical data representation. Token annotation properties pointing to literal values on Wikibase can be straightforwardly declared equivalent to the NIF ontology (Hellmann et al., 2013), and lexical data on Wikibase is modelled following Ontolex-Lemon (McCrae et al., 2017). On the other hand, we are not aware of established standards regarding the interface between the corpus and the lexicon, and, more precisely, regarding properties that link text tokens and token spans to dictionary entries at *lexeme* (lemma, headword), sense, and form level, that is, in all these cases and unlike according to NIF, linking not to literals but to entities (nodes on the graph).

As for the software solution to deploy for the implementation of the model and subsequent population with data, we choose Wikibase, which is the only free graph database software we are aware of that features user management, editable entity data exhibition pages, reversible edit histories, and graphical and grammatical SPARQL query endpoints.

The use case for which we propose this model is documents that belong to the Basque Historical Corpus, although given the general character of the model we claim that it can serve in other contexts, too⁵. That corpus contains literature written in Basque from before 1900, and today it exists in the different versions referenced above, and stored in separated and incompatible data silos (based on relational databases) and made available through different online user front ends. In this presentation, we focus on the corpus-lexicon interface, while our model also should be as flexible as to allow associating source text tokens with their representation in different editions of the source⁶, so that different versions of the text could be displayed, together with all kinds of annotations, using the dataset stored on Wikibase.

⁵ A recent experiment following the proposed model and involving Serbian corpus data modelled according to NIF and an Ontolex dictionary is documented at <https://serbian.wikibase.cloud>.

⁶ We think of a property pointing from a token or token span to a string representation of the token, qualified with a property pointing to the corresponding edition.

3. Data Model and Implementation

Heavily inspired by the latest trends in the field of Linguistic Linked Open Data⁷, we model a corpus token as node in a knowledge graph, and link it to the following data objects (cf. Figure 1):

1. the respective paragraph in the source document (that is, for example, a transcription carried out and stored on the Wikisource platform);
2. a lexeme node, which is annotated with the standard Basque lemma (we call this property *has standard lexeme*);
3. a lexical form associated to that lexeme, which is annotated with the grammatical features describing the form (*has standard form*);

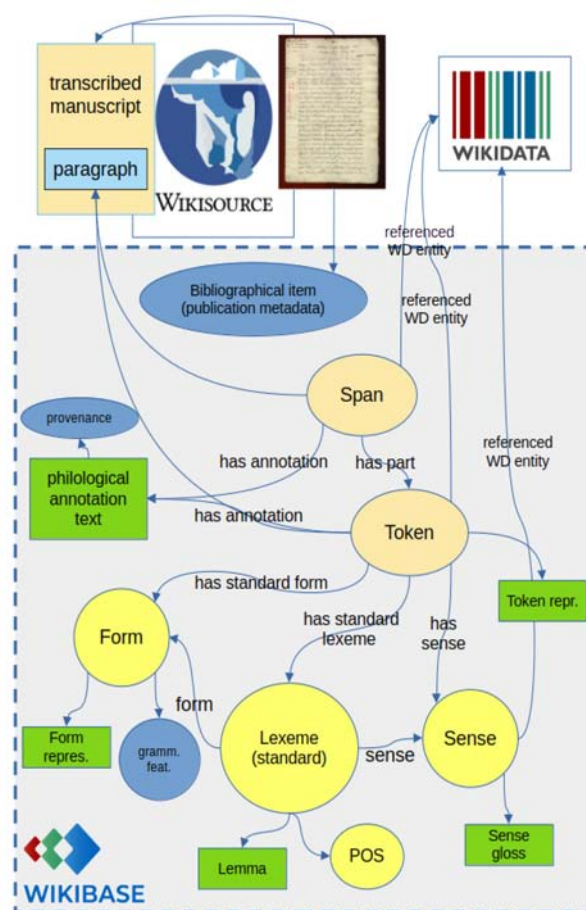


Fig. 1: Proposed model for the corpus-lexicon interface

4. a lexical sense associated to that lexeme, which is annotated with a sense gloss (we name it *has sense*, since here the distinction is not standard vs. non-standard but whether a sense is to be flagged as historical (i.e., in use in some past period));

⁷ For recent related work, see Stanković et al. (2023).

5. an ontology concept that matches to the word sense. Note that the redundancy in the properties named *referenced WD entity*, with both token and dictionary sense as subject, is intended: There will be cases where a dictionary sense is already aligned to an ontology concept⁸, and there will be others, where automatic or human annotators link a text token to a named entity, which is referred to using a Wikidata item ID. Values of this property attached to either the token or to the dictionary sense can be inferred from each other.
6. to a text chain containing philological annotations, such as, in our use case: *instead of something crossed out, as side note, hardly decipherable*, etc.

Furthermore, we represent token spans as separate nodes; these are linked to the contained tokens, and to annotations that apply to the whole span.

We implement and populate the model on our own Wikibase instance⁹ hosted on Wikibase Cloud¹⁰. We chose that software for a range of reasons:

1. The software itself and the chosen hosting solution are free and open-source software, with an active community dealing with its development and user support.
2. Uploaded contents are instantly exhibited in a graphical user interface which allows for collaborative editing of every single claim made on the corpus token, including reversible edit histories.
3. The created dataset is publicly queryable using SPARQL.
4. We are able to export the dataset as an RDF/OWL document, as far as classes and properties on the Wikibase are aligned with external RDF vocabularies¹¹.
5. The dataset is compatible with Wikidata, in the sense that lexemes, forms, senses, their descriptions, and ontology concepts can be aligned between our own Wikibase and Wikidata, which in the Wikibase Ecosystem is called *federation*¹², so that it is possible to launch queries that deliver relations from both graphs at the same time¹³.
6. Core classes and properties used on a Wikibase by default for describing lexemes deploy Ontolex-Lemon, the W3C-recommended model for lexical data, so that the created datasets are compatible with Wikidata and other resources deploying that model; a relation between entities of the same classes in two different datasets or Wikibase instances can be straightforwardly set.

⁸ According to Ontolex-Lemon, that link is named *ontolex:reference*. On Wikidata, it is <https://www.wikidata.org/entity/P5137>.

⁹ Accessible at <https://monumenta.wikibase.cloud>.

¹⁰ Wikibase Cloud is a service provided by Wikimedia Deutschland, see <https://wikibase.cloud>.

¹¹ See a Wikibase property declared equivalent to a NIF property at <https://monumenta.wikibase.cloud/wiki/Property:P147#P179>.

¹² See <https://meta.wikimedia.org/wiki/LinkedOpenData/Strategy2021/Wikibase>.

¹³ For example, a standard Basque lexeme on Wikidata is linked to additional sources where it is described; Wikidata also contains further knowledge about named entities such as birth dates for persons or geographic coordinates for locations.

4. Population of the Model: A First Use Case

We are currently populating the proposed model with tokens from a 1737 Basque manuscript, the transcription of which has been carried out on Wikisource¹⁴, and inserting annotations of the above described types including philological annotations by Lakarra (1985), and own annotations. The Wikibase items created to describe each token link back to the corresponding paragraph in the manuscript transcription on Wikisource, which is on that platform accessible together with digital images of the source manuscript.

All annotation operations, including those pointing to dictionary entries, forms, or senses, can be set manually in the Wikibase interface. In order to shortcut the operations, we have defined functions in a script that responds to the user via a web application¹⁵. Tasks addressed by these functions are token entity creation, token span definition, dictionary linking, and text annotation display.

5. Summary

We have proposed a model for a corpus-lexicon interface, implemented it on a Wikibase instance, and started to populate it with a Wikisource transcription as the first textual source. We also have produced scripts that ease the different workflow steps, more precisely, a web application that allows a user to define token spans, and to add annotations the model allows for on token or span level. We want to point out the advantages of having chosen Wikibase as working platform: All tokens and token spans have their own entity data page online, where annotations can be reviewed and edited. All edits are recorded in the edit history of each token or token span item, on the one hand, and in the user contribution records, on the other. That means the platform is ready for collaborative work on the data, e.g., in a larger project where different human editors are involved. Through the Wikibase API, annotations can also be set massively; this is useful when results stemming for example from a lemmatization or Named Entity Recognition tool shall be uploaded and then validated by human annotators.

Through SPARQL, advanced queries are enabled; these may involve all token or span annotations, their context, the source document metadata, and everything that can be inferred from the Wikidata alignments at different levels¹⁶.

References

Bilbao, G., Gómez, R., Lakarra, J. A., Manterola, J., Monoule, C., & Urgell, B. (2011). *Lazarraga eskuizkribuaren edizioa eta azterketa*. Lazarraga Eskuizkribuaren Edizioa Eta Azterketa, Vitoria-Gasteiz: UPV-EHU. <https://www.ehu.eus/monumenta/lazarraga/>

¹⁴ Accessible at https://eu.wikisource.org/wiki/Azkoitiko_Sermoia.

¹⁵ We are using a python Flask webapp, which calls functions involving the wikibaseintegrator python module for communication with Wikibase. See <https://monumenta.wikibase.cloud/wiki/Project:About> for reference of the code repositories.

¹⁶ See https://monumenta.wikibase.cloud/wiki/Larramendi,_Azkoitiko_Sermoia.

Estarrona, A., Etxeberria, I., Soraluze, A., Etxepare, R., & Padilla-Moyano, M. (2022). The first annotated corpus of historical Basque. *Digital Scholarship in the Humanities*, 37(2), 391–404. <https://doi.org/10.1093/llc/fqab066>

Hellmann, S., Lehmann, J., Auer, S., & Brümmer, M. (2013). Integrating NLP Using Linked Data. In H. Alani, L. Kagal, A. Fokoue, P. Groth, C. Biemann, J. X. Parreira, L. Aroyo, N. Noy, C. Welty, & K. Janowicz (Eds.), *The Semantic Web – ISWC 2013* (pp. 98–113). Springer. https://doi.org/10.1007/978-3-642-41338-4_7

Lakarra, J. A. (1985). Literatur gipuzkerarantz: Larramendiren Azkoitiko sermoia (1737). *Anuario del Seminario de Filología Vasca ‘Julio de Urquijo’*, 19(1), 235–281. <https://doi.org/10.1387/asju.7685>

Larramendi, M. (1745). *Diccionario trilingüe castellano, bascuence y latin dedicado a la M.N. y M.L. provincia de Guipuzcoa*. Bartholomé Riesgo y Montero. <https://www.kmliburutegia.eus/Record/203133>

Lindemann, D., & Alonso, M. (2021). A workflow for historical dictionary digitisation: Larramendi’s Trilingual Dictionary. *Electronic Lexicography in the 21st Century: Proceedings of the eLex 2021 Conference*, (pp. 598–614). https://elex.link/elex2021/wp-content/uploads/2021/08/eLex_2021_39_pp598-614.pdf

McCrae, J., Bosque-Gil, J., Gracia, J., Buitelaar, P., & Cimiano, P. (2017). The OntoLex-Lemon Model: Development and Applications. In I. Kosem, C. Tiberius, M. Jakubíček, J. Kallas, S. Krek, & V. Baisa (Eds.), *Electronic lexicography in the 21st century: Lexicography from scratch. Proceedings of eLex 2017* (pp. 587–597). Lexical Computing CZ s.r.o. <https://elex.link/elex2017/wp-content/uploads/2017/09/paper36.pdf>

Stanković, R., Chiarcos, C., Utvić, M., & Kitanović, O. (2023). Towards ELTeC-LLOD: European Literary Text Collection Linguistic Linked Open Data. In S. Carvalho, A. F. Khan, A. O. Anić, B. Spahiu, J. Gracia, J. P. McCrae, D. Gromann, B. Heinisch, & A. Salgado (Eds.), *Proceedings of the 4th Conference on Language, Data and Knowledge* (pp. 180–191). NOVA CLUNL, Portugal. <https://aclanthology.org/2023.ldk-1.16>

Vrandečić, D., & Krötzsch, M. (2014). Wikidata: A Free Collaborative Knowledge Base. *Communications of the ACM*, 57, 78–85. <https://doi.org/10.1145/2629489>

Acknowledgements

The research leading to the presented results has been supported by DLTB research group (Basque Government, IT1534-2) and Monumenta Linguae Vasconum research project (MINECO, PID2020-118445GB-I00, Govt. of Spain).

Contact Information

David Lindemann

UPV/EHU University of the Basque Country
david.lindemann@ehu.eus

Mikel Alonso Arrospide

UPV/EHU University of the Basque Country
mikel.alonsoa@ehu.eus

Ivana Lovrić Jović and Martina Kramarić

THE DUBROVNIK IDIOM THROUGH TIME

Crafting a Diachronic Dictionary

Abstract This article discusses the project, Dictionary of the Dubrovnik Idiom, conducted at the Institute for the Croatian Language. The project aims to develop a born-digital diachronic dictionary of the Dubrovnik idiom, covering the period from the 16th century to the end of the 20th century. The dictionary will be based on a historical corpus compiled within the project's scope, featuring texts from the same period. Upon completion, this dictionary will be publicly available in digital form, providing valuable insights into the linguistic evolution of the Dubrovnik region. The creation of such a dictionary will meet the needs of the scientific and cultural public as well as the citizens of Dubrovnik. Specifically, the Dubrovnik idiom is relevant because it played an important role in the standardization of the Croatian language. Additionally, well-known Renaissance and Baroque literary works are composed in it, and today it is in decline, which concerns its speakers.

Keywords born-digital dictionary; diachronic dictionary; historical corpus; Dubrovnik idiom; historical lexicography

1. Introduction

The creation of a diachronic dictionary that presents a language before its standardization poses special challenges to lexicographers, who need to rely solely on written sources or previous scientific descriptions of that language. If the dictionary presents a vernacular language,¹ the challenges are even greater, requiring a methodology that combines historical linguistics and dialectology approaches. In the newly started project, “Dictionary of the Dubrovnik Idiom”, we are facing various challenges, which we address in this article

2. Background and the Reasons for the Creation of a Dictionary of Dubrovnik Idiom

2.1 The Value of Dubrovnik Idiom in the Croatian Philology

In the standardization of the Croatian language, the Dubrovnik literacy of the 16th and 17th centuries, based on the Dubrovnik idiom, contributed significantly to establishing the Štokavian dialect as the foundation for the Croatian Standard language. This fact is well known in the history of the Croatian language, and the language of Dubrovnik literacy was analyzed and described quite early (Rešetar 1933, 1941). The lexical

¹ In this article, we are using vernacular (colloquial) language as the opposite term to standard language; i. e., the language that is spoken among locals and is not standardized. Furthermore, this dichotomy is used to distinguish between the literary high-register language, referred to as the Dubrovnik language, and the folk colloquial language, referred to as the Dubrovnik vernacular. This dichotomy was introduced by Lovrić Jović (2006; 2014, pp. 6–14).

component of Dubrovnik literacy is well represented in the Academic Dictionary of the Croatian or Serbian Language (ARJ 1880–1976), an extensive historical dictionary of the (not only) Croatian language from the 12th to the 20th century, written in 23 volumes. Due to its extensiveness, it is challenging for the general public to access and find specific lexemes that belong exclusively to Dubrovnik literacy.

Some researchers have also included descriptions of the local vernacular of the 19th century as part of their dialectology studies of the Croatian language (Budmani, 1883). Further analysis of the local vernacular of the 18th century was conducted by Lovrić Jović (2014), who examined its traces in the literary language of *Francesarie* – Molière’s comedies adapted into the Dubrovnik idiom. This analysis reveals significant features of the local vernacular of that time. Similar conclusions can be drawn for the administrative language of testimonies, which were transcribed and analyzed by the same author (Lovrić Jović, 2006, pp. 173–192; 2015).²

2.2 The History of the Dubrovnik Idiom

The local vernacular of Dubrovnik originally belonged to the Štokavian dialect group, which is part of the South Slavic branch of the Slavic language family. Starting in the 16th century, it began to evolve and incorporate features of the newer Štokavian dialect (Vulić, 2011, pp. 180–182). This process was completed in the 19th century, so today’s Dubrovnik speech is undoubtedly New Štokavian with Ijekavian pronunciation.³ According to Brozović (2005, p. 20), the process of “new Štokavianization” in the 17th and 18th centuries was performed inconsistently, resulting in special, authentic features of the idiom, particularly in terms of prosody.

Throughout its history, the Dubrovnik idiom was heavily influenced by the Italian language. This influence affected the local vernacular, resulting in the incorporation of many Italianisms. Consequently, a significant number of these Italianisms are expected to be part of the word list in the Dictionary of the Dubrovnik Idiom.⁴

As for the Italian dialect that influenced the Dubrovnik idiom, according to Muljačić (1962, p. 341), the Venetian language initially had the greatest influence, but was later supplanted by Tuscan, which the Dubrovnik residents learned from books or in Italy.⁵ The local citizens were less familiar with Tuscan. The use of Venetian declined among merchants and sailors, while Italian, supplanted Venetian as the cultural language learned at Italian universities and by merchants.

² In the project The Dictionary of Dubrovnik Idiom and in the process of its creation of the corpus and dictionary, we will emphasize the difference between high and low register literacy. The special attention will be given to the low register literacy (to be explained) because it represents the local Dubrovnik vernacular in its best manner, and its lexis, albeit stylized, offers the possibility of presentation the local vernacular in the diachronic perspective from the 16th till 20th century.

³ This distinguishes it from other Štokavian dialects like the Ekavian and Ikavian varieties.

⁴ This complex sociolinguistic situation was described throughout the centuries by Lovrić Jović, 2006, pp. 173–175. Italianisms in the Croatian language were analyzed by Muljačić, 1973; Sočanac, 2004; and Lovrić Jović, 2006.

⁵ That language exerted no direct influence, it was acquired from books or introduced by the considerable number of Dubrovnik students who pursued studies in Italy.

2.3 Dubrovnik Idiom in Dictionaries

Despite the significant role that the Dubrovnik idiom played in the standardization of the Croatian language, there is still no comprehensive dictionary of the Dubrovnik idiom based on a corpus and adhering to rigorous lexicographical principles. Instead, recent decades have seen the creation of glossaries by speakers of the Dubrovnik vernacular. These glossaries lack in quality and expertise. However, there does exist a scientific foundation-based, non-corpus-based Dubrovnik idiom dictionary for the 20th century (Bojanić & Trivunac, 2002). For the older linguistic layer, there is the dictionary by Jakov Mikalja (1649/51, reprint Mikalja, 2011), which is based on the Dubrovnik idiom and is considered a precursor to the modern type of dictionary. To this, we can add the annex dictionaries that belong to the monographs of the Dubrovnik idiom (Lovrić Jović, 2014, 2015; Lovrić Jović & Jozić, 2016).

3. The State of the Croatian e-Lexicography

The ‘poor’ state of Croatian e-lexicography was thoroughly described by Despot & Möhrs (2015, pp. 334–335; pp. 345–346). The situation underwent significant enhancement with the implementation of the project “Croatian Web Dictionary *Mrežnik*” (from 2017) and the initiation of the development of the first corpus-based, born-digital contemporary monolingual e-dictionary of contemporary Standard Croatian, *Croatian Web Dictionary – Mrežnik*.⁶ The dictionary is currently in the process of creation, with only a demo version available, which covers entries to the letter F (<https://rjecnik.hr/mreznik/>). Comprising three modules, the dictionary heavily emphasizes the usage perspective of language, particularly focusing on collocations. Hence, the absence of a historical and dialectal e-dictionary in Croatian philology persists. Our project endeavors to address and fill this void.

4. The Project of a Dictionary of Dubrovnik Idiom

Recognizing the scientific and general need for an integrated dictionary, a project of compiling a dictionary of the Dubrovnik idiom has commenced at the Institute for the Croatian Language (from 2024).⁷ This project aims to create a corpus-based born-digital diachronic dictionary of the Dubrovnik idiom covering the period from the 16th till the end of the 20th century.⁸ As this dictionary is intended to fill a significant gap, it should serve several user groups. First of all, it would be aimed at language historians and dialectologists, with the possibility of extending it to theatre scholars,

⁶ The project is being conducted at the Institute for the Croatian Language and is led by Dr. Lana Hudeček, who is a as well member of the Dictionary of Dubrovnik Idiom project. Dr. Hudeček’s extensive lexicographical knowledge and expertise will be inherited by the team involved in this project.

⁷ Project is led by Dr. Ivana Lovrić Jović.

⁸ A similar concept to the dictionary project can be found in the printed dictionary by Bencistà, Alessandro (2021). *Vocabolario del vernacolo fiorentino toscano (con gli esempi delle voci da Dante ai moderni)*. This dictionary presents the Italian Florentine dialect from the 14th century, which was selected as the foundation for Standard Italian language owing to Dante’s literary works. In this dictionary, examples from both the vernacular and literary language are integrated, and a diachronic approach is employed. The dictionary contains a total of 400 entries.

dramaturges, and actors. Another important target group would be the citizens of Dubrovnik, especially those who are concerned about the possible disappearance of the Dubrovnik vernacular.⁹

4.1 Terminological Clarifications Regarding Dubrovnik Idiom

We are using “idiom” as the superordinate term, which incorporates both the local Dubrovnik vernacular, both in the form of the local speech and in the form of the stylized literary language presented mainly in low-register literature (comedies, congratulatory folk songs, satirical periodicals, etc.), and the literary language written primarily in high-register literature. Since the vernacular, as an authentic living expression, is impossible to explore for past times, information about it can only be obtained indirectly from those parts of written texts where the Dubrovnik vernacular is best reflected (dramatic works, congratulatory poems, transcribed handwritten testimonies)¹⁰ and in genres that effectively convey spoken language, for example, in epistolary prose. For more recent periods, the lexicon will be excerpted from audio recordings of speakers of the Dubrovnik vernacular, among the most valuable of whom are emigrants to other continents who best preserve the speech of the second half of the 20th century.¹¹

5. Material for the Corpus

The Dictionary of the Dubrovnik Idiom will be based on a historical corpus of texts from the same period as the dictionary. It will be compiled within the scope of the project and built from scratch. In selecting material for inclusion in the corpus, emphasis is placed on the diversity of genres, functional styles, and temporal coverage.

The corpus is compiled from various sources, including different historical texts that represent various literary and functional styles of the language. It will contain existing philological descriptions of the Dubrovnik idiom: grammar, glossaries, dictionaries. Further it will include administrative-legal texts such as testimonies from the 17th and 18th centuries (Lovrić Jović, 2015), archival documents like *Lamenta Ragusina* and *Lamenta del Criminale* from the 17th and 18th centuries, personal letters and epistolary prose from the 19th century, folk congratulatory songs, literal and dramatic texts such as Držić’s comedies and *Francesarie* – adaptations of Molière’s comedies into the Dubrovnik idiom written in the 18th century (22 comedies, Lovrić Jović, 2014), memoir texts of prominent citizens, published chronicles, etc. Further, we will incorporate sources from a special section of Dubrovnik press from the 20th century (including satirical newspapers published from 1901 until the end of the 20th century), and a cookbook representing Dubrovnik dishes throughout history and written in the Dubrovnik vernacular (Ničetić, 2012).

⁹ The phonological, morphological, syntactical, and prosodic features that rendered it authentic are nearly extinct.

¹⁰ Deeper theoretical explanation of this terminology see in Lovrić Jović, 2006, pp. 173–175; 2014, pp. 6–9.

¹¹ Audio recordings were recorded ten years ago among local inhabitants of the city of Dubrovnik by Ivana Lovrić Jović.

In addition, we will use as a reference source the previously mentioned Dictionary of Mikalja and the ARJ, in which the canonical high literature was processed and lexical material of prominent Dubrovnik writers through the centuries was mentioned (Đore Držić, Šiško Menčetić, Mavro Vetranović, Marin Držić, Ivan Gundulić, etc.), along with early testimonies of literary works such as prayer books, lectionaries and miscellanies. As an additional source for the dictionary, we will include audio recordings of speakers from Dubrovnik.

Once the corpus is completed, it will be representative in terms of covering different time periods and functional styles of language. Both stylized vernacular, colloquial low-register vocabulary from comedies, congratulatory songs, satirical periodicals, etc., as well as literal high-register vocabulary, will be presented. The representation of various sources from different time periods will enable the analysis of lexicon in a diachronic overview and facilitate the recognition of changes that have occurred over time in the lexicon (disappearance, expansion, or narrowing of meanings, changes in forms, etc.).

Currently, the project is in the phase of creating the corpus using the Sketch Engine, which involves the digitization and critical interpretation of written sources from various centuries. The corpus is derived from printed documents, manuscripts, and TIFF image files. All the documents will be digitized for the first time within the scope of the project. In addition to digitized sources, handwritten manuscripts such as personal letters from the 19th century will not be transcribed fully, but they will be consulted as sources for examples.¹² The same approach will be applied to audio recordings, which will not be fully transcribed. Instead, only examples will be excerpted. All excerpted material will be adapted according to contemporary transcription rules, and all data in the dictionary will be provided in the transcribed form.¹³ For consistency, some words from printed material will be adapted according to the transcription principles adopted for this dictionary. This is considering the fact that printed texts may contain forms written according to different transcription principles. Once the corpus is finished, it will be searchable, upgradable, and partly publicly available, with exceptions for works that are still restricted by legal rights.¹⁴

¹² For sources written in manuscripts during a time when there was no Standard Croatian orthography yet, and Italian orthography was used on the coast, especially for consonants with diacritics (such as *č, ć, ž, š, dž, đ*, but also *lj* and *nj*), transcribing all the personal letters in full would be very time-consuming.

¹³ The same rules were used, for instance, in the transcription of testimonies by Lovrić Jović, 2015. In addition to the transcribed texts, the corpus may also contain non-transcribed texts written in the old, non-standardized orthography. These may be digitized printed material from the 19th century or older, and later in the text we will explain how we will extract data for the dictionary from these texts. If the user of the corpus wants to search for an element in the corpus, they need to know how this element might have been written down over the centuries.

¹⁴ Currently, the corpus is only accessible to members of the project team.

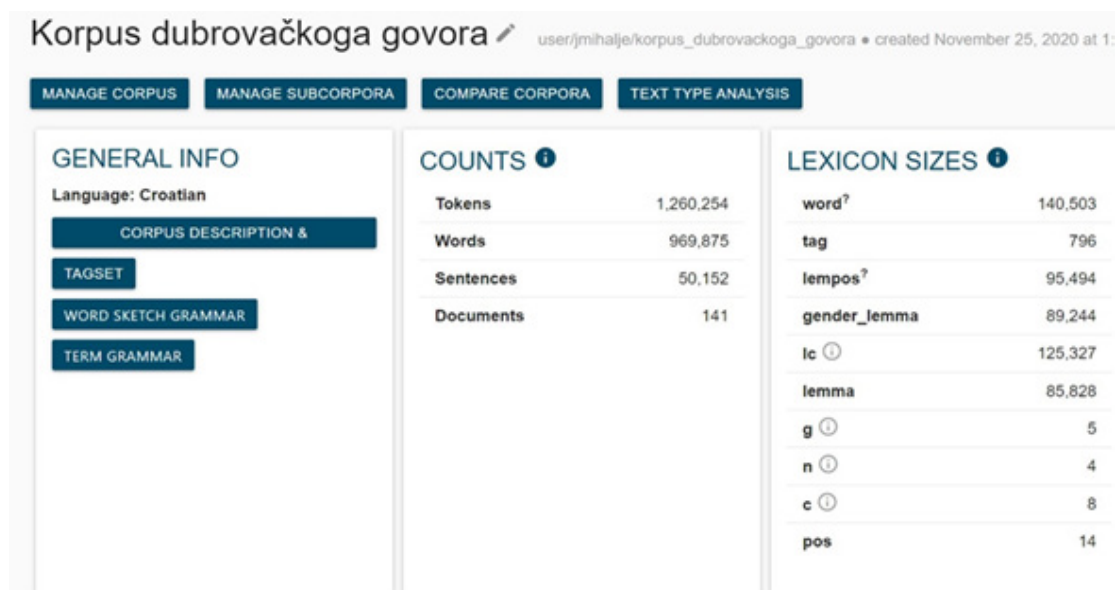


Fig. 1: Current status of the corpus

6. Methodology

In crafting the Dictionary of Dubrovnik Idiom, we will integrate methodologies from both historical linguistics and contemporary dialectology dictionaries. Once the dictionary is finished, it will provide crucial information for understanding the diachronic development of the described language. Additionally, it will offer valuable insights for dialectologists in terms of accentuation and sound recording of the headwords.

As mentioned before, the methodology of the dictionary follows the *Mrežnik* dictionary with the exception that only one module will be created. This means that a corpus-based methodology will be employed.¹⁵ The microstructure of the dictionary entry will be similar to the mentioned dictionary but with exceptions due to historical component of the dictionary and the limitations of the corpus.¹⁶ Therefore, only very typical collocations and phrases will be provided, in contrast to *Mrežnik*, where collocations play an important role (Hudeček – Mihaljević, 2020b, p. 78). T-Lex was chosen as the DWS for both projects. The microstructure of the dictionary entry is shown in the diagram below.¹⁷

¹⁵ In the *Mrežnik* project, lexicographers use a corpus, but they have the freedom to decide what should be included in the dictionary entry. All data extracted from the corpus serve only as guidelines (www.mreznik.ihjj; Hudeček – Mihaljević, 2020a, p. 657).

¹⁶ The second major difference concerns the fact that *Mrežnik* is the normative and the Dictionary of the Dubrovnik Idiom the descriptive dictionary. About the normative aspect of *Mrežnik* see in Hudeček & Mihaljević 2020a, pp. 658–659.

¹⁷ The microstructure of the dictionary entry was created by Ivana Lovrić Jović and Lana Hudeček.

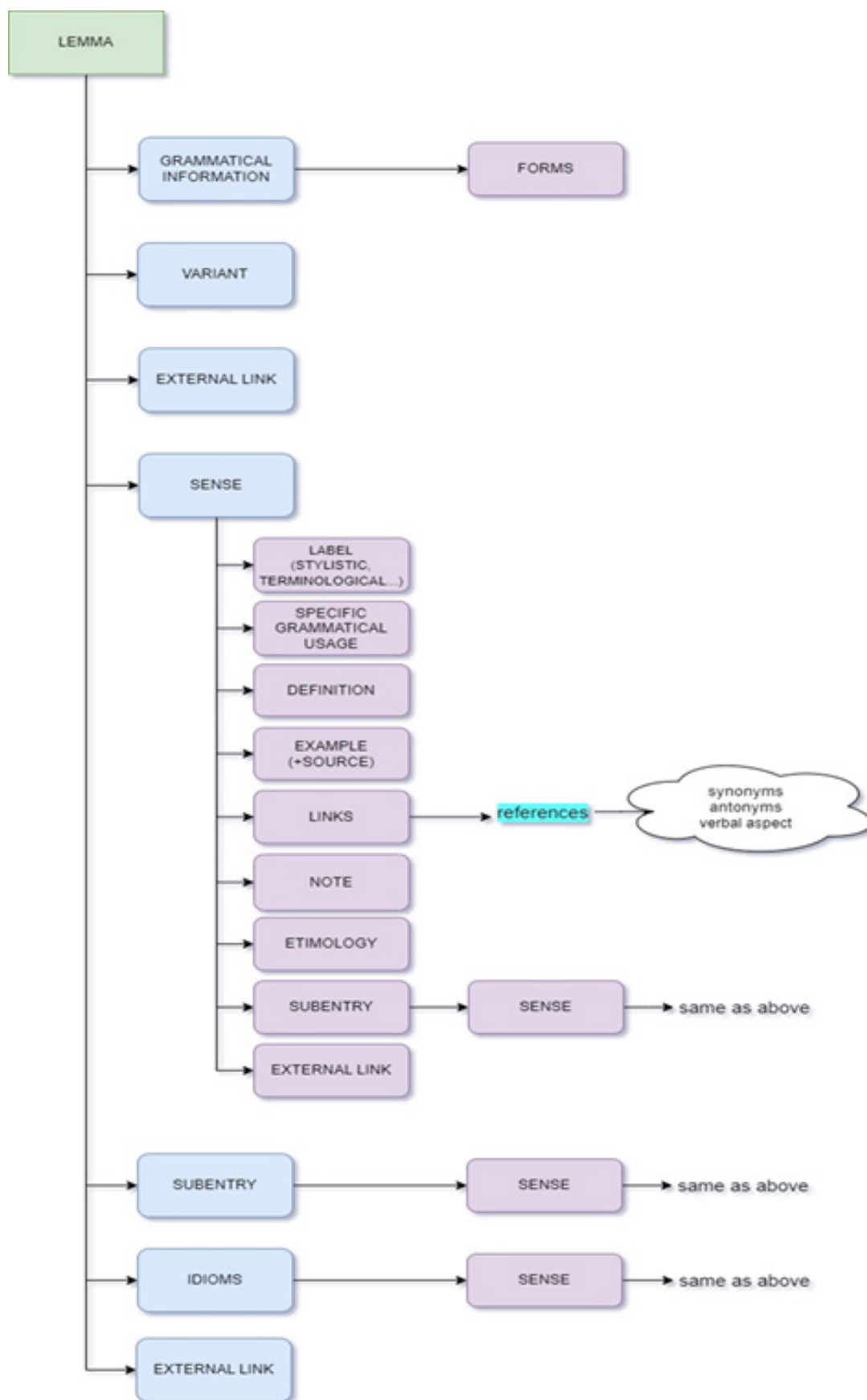


Fig. 2: The microstructure of the dictionary entry

The corpus of the Dubrovnik idiom is lemmatized, and we will use the Sketch Engine web tool, but the extraction from the corpus will be performed mainly manually.¹⁸ As the additional tool, we will use *Word Sketches* for Croatian which were written for the *Mrežnik* project (Hudeček & Mihaljević, 2020a, p. 654).¹⁹ Since the Dubrovnik idiom is based on the same Štokavian dialect as standard Croatian, we will be able in some extent to display lexeme context through *Word Sketches*, even if we are working with historical lexemes that sometimes have a different form than contemporary lexis.

6.1 Word List

Following the principle of effectiveness (Morkovkin, 1996), the dictionary will be distinctive in a broader dialectal sense, adhering to the concept of differentiating vocabulary where items “that are equal to the standard” are not included (in the very broad sense of the word). This means that, since this isn’t a standard language dictionary, only those lexemes that are unfamiliar to other speakers of the Croatian language and that constitute the general lexicon of Dubrovnik residents (not specialized) will be excerpted. Words that differ only in their phonological and morphological form from their standard equivalents will not be processed in this version. Professional vocabulary will not be considered, or rather, it will be considered only to the extent that it is usually represented in standard language dictionaries. Lexemes belonging to the standard language (of which there are many in every idiom, e.g., *water, head, house, child, sky, live, know, wait*, etc.) will not be included.²⁰

Initially containing around 730 entries, the dictionary will be open to further expansion. Since the project is currently in the corpus creation phase and the lexicographic processing of headwords has not yet begun, we will present each component of the envisioned dictionary entry and outline the challenges expected in the processing considering the limited corpus and lack of all confirmations, as well as the variability of forms, which is an important characteristic of historical and non-standard language.

¹⁸ In view of the fact that there is not yet a tool that can cope with the non-standardised spelling in the corpus, we have to carry out the search manually. Since the members of the project team are philologists and historical linguists, they all know the possible spellings of a lemma in the old orthography. For example, *mobilja* ‘furniture’ is written as *mobiglia*, *crkva* ‘church’ as *zarqua*, *poći* ‘to go’ as *pochi*. The old orthography differs from the standard orthography only in certain letters and is well known to the project members. Once the data written in non-standard orthography is extracted from the corpus, the headwords and examples of usage will be transcribed for the dictionary entry.

¹⁹ All corrections in the forms of headwords will be performed manually. The display of lexeme context through *Word Sketches* enables an overview of the most common collocations sorted into syntactic categories, facilitating the discovery of good examples of word usage or collocations (Hudeček & Mihaljević, 2020a, p. 654).

²⁰ Furthermore, the decision on word lists for the dictionary cannot be based solely on the corpus, frequency of occurrence of an element, or collocations, but rather on experience from which it is possible to conclude whether a lexeme is sufficiently confirmed in the Dubrovnik idiom to deserve a place in the dictionary. Namely, there are lexemes appearing in a part of the corpus that will not enter the dictionary because they are confirmed only once, unknown to speakers, and based on the genre of the text, it can be concluded that they are failed neologisms. The majority of them come from various folk texts, from congratulatory poems to satirical periodicals, but also in Dubrovnik adaptations of Molière’s comedies.

7. Dictionary Entry

7.1 Headword: Form, Accentuation, Variants

The headword will be accentuated and given in the canonical form (e.g., for nouns: nominative singular, for verbs: infinitive form), and in the form transcribed according to contemporary transcription rules.

Headwords that are not neutral will be labeled with stylistic and terminological labels.

In non-standard dictionaries, it is anticipated to encounter phonological variants of the headwords (e.g., *grakàtež/grahatež/grakatež* for ‘grater’), which in this dictionary will all be listed. These variations offer us a crucial piece of information about the development of the language, which is an important part of the diachronic perspective of this dictionary. For example, in the word *pacijencija/pacijenca* (‘patience’), phonological variations exist, and the first version is older. Similarly, for the word *gòspoja/gospòda*, phonological variations exist, and the first form is older and belongs to literary language. In the verb *afiktat(i)/afitat(i)* (‘to rent’), both variants are recorded in the same period, in the same testimony from 1767, but the first version (with the consonant cluster *-kt-* indicating an older (Latin) layer of loanword entry into the Dubrovnik vernacular) has completely disappeared to this day, and the second one is still in use. This information will be provided through stylistic labels followed by explanatory notes as part of the dictionary entry.

If the canonical form of the word is not attested in the corpus, it will be reconstructed on the basis of the phonological and morphological features of the Dubrovnik idiom. If there are two or more possible forms for this construction, all of them will be listed. The headword will be accentuated, and a sound recording will be provided.²¹ As the corpus generally does not offer information about accents, deciding whether to record them presents a significant challenge.²² It will have to be sought in the scientific lexicographic literature.²³ Indirectly, information will be sourced from scientific phonological descriptions of the older Dubrovnik language in literature, from dramatic texts in which the accents of Dubrovnik actors born at the beginning of the 20th century, etc. In addition, the aforementioned audio recordings of Dubrovnik speakers will be used for accent reconstruction. In cases where the accent is disputed, it will be reconstructed based on living speakers. The sound recording of the headword will be captured in a music studio and pronounced by the project leader, with collaboration from a Dubrovnik actor possessing extensive experience in portraying Dubrovnik characters in dramas by Dubrovnik authors spanning from the 16th to the 20th century.

²¹ The accentuation of the headword in the dictionary will be performed by a member of the dialectology team, with suggestions from the project leader.

²² For example, we will provide an overview of the accent reconstruction for the headword *kondicion* (‘condition’), an Italian loanword. In the Dubrovnik idiom and in the Standard Croatian language, the stress behaves differently. In the standard language, there is a shift in stress (from the original *condizione*) by one syllable closer to the beginning of the word. In the Dubrovnik idiom, the stress from the Italian origin, which falls on a diphthong with the first element being a semivowel /j/, shifts directly to that semivowel, which becomes the bearer of stress: Italian *kondit’tsjo* – Dubrovnik *kondiciōn*. In the standard language, the stress moves one syllable further from the end of the word: *kondicija*, or according to the Neoštokavian accentual system, even further: *kòndicija*.

²³ In ARJ headword is accentuated.

7.2 Grammar Block

A grammar block will contain information about the morphological features of the word, the part of speech, gender, and usually for nouns, the form of the genitive singular case, and for verbs, the aspect information will be given.²⁴ For adjectives, we will provide information about the definitive and indefinite forms. Exceptions to that rule will be mentioned as specific grammatical usage if an unexpected form of some word occurs.²⁵

If the grammatical information is not attested in the corpus, some forms will need to be reconstructed according to available insights from phonological and morphological descriptions – e.g., *pantarula* Gsg, Nsg *pantaru* (not *pantarul*) ‘fork’.²⁶

7.3 Sense Block

A sense block will contain definitions covering all meanings and examples of usage based on the corpus. The definitions will be conveyed either through a Croatian standard synonym (e.g., *vizita* – *posjet* for ‘visit’, *kondicion* – *uvjet* for ‘condition’) or explained via paraphrase (e.g., *pižuo* – ‘specific kind of stone’). Examples of usage will be drawn from the corpus based on their degree of clarity, with the explanatory note indicating from which text it was excerpted and the time period and genre/functional style. This will enable users of the dictionary to determine whether a word belongs more to colloquial, low-register language or literal, high-register language.

In the absence of examples of usage from the corpus or if they are not sufficiently explanatory, they may be found in philological works, ARJ, or from existing dictionaries (with a critical approach), or from data provided by living speakers. Typical semantic relations, like synonymy and antonymy, between words will be recorded. For synonymy relations, we will mention only synonyms belonging to the Dubrovnik idiom because we already used a synonym from the standard language for the definition. So, for the lexeme *parac* (‘patron saint’), the Dubrovnik synonym will be *protektur* and the standard definition *svetac zaštitnik*. For the lexeme *pedepsa* (‘punishment’), the synonym will be *kastig* and the standard definition *kazna*.

²⁴ Once more we will mention the word *kondicion*, which exhibits many specificities of the Dubrovnik idiom on several levels: phonological, morphological, and prosodic. Regarding its grammatical gender specificity compared to the standard language, its ending *-(i)on* can be used for feminine nouns in Dubrovnik, while in the standard language, they are masculine nouns (*bastion*, *peron*, *žeton*). The noun *kondicion* originates from the Italian form *condizione*, more likely the Venetian form *condizion*, and it is one of many borrowed into the Dubrovnik idiom with this ending or rarely its phonological variant *-ijun* (*okupacion*, *okazion*, *ezekucion*, *ordinacion*, *komisijun*). In the standard language, loanwords of this type adapt on a derivational level and take the ending *-ija* (*kaucija*, from German *Kaution* ← Latin *caution*; *institucija* from Latin *institutio*; *špedicija* from Italian *spedizione*). In the Dubrovnik idiom, such loanwords either drop the initial vowel *-e* (from Latin or Italian origin) or they enter the language without it (from Venetian origin).

²⁵ For example, regarding the word *kondicion*, in addition to the genitive form *kondicioni*, we will also discuss the instrumental form, which is found several times in testimonies from the 17th and 18th centuries, confirming the old case ending for the instrumental singular: *kondicionim*. Alongside the archaic form, it is necessary to mention the contemporary form of the instrumental singular, as it is confirmed in the same texts: *kondicioni*.

²⁶ The expected form of the nominative singular *pantarul* according to the genitive singular *pantarula* does not belong to the Dubrovnik idiom because the vocalization *l* to *o* appears in the nominative singular. That change is typical for the Štokavian dialect.

In addition, for the verbs, we will provide both the perfective and imperfective forms. It is common for perfective verbs to be formed from basic verbs by adding prefixes, and these prefixes can vary. Therefore, for the Latinism *kordati* ('to arrange, set in order, organize, agree, come to terms'), two versions of the perfective aspect will be listed: *uskordati* and *akordati*.

When the corpus allows, we will mention collocations or typical phrases.

7.4 Etymology

For loanwords, an etymological note will be provided. In the Dubrovnik idiom, aside from Italianisms, these include Latinisms, Hellenisms, Turkisms, and occasionally rare Germanisms, Hungarianisms, and Hispanicisms. Italianisms will sometimes be specified as Venetianisms if the information is certain (if it's probable but uncertain, that will also be noted).

The note section will be used for various types of notes, including different variants of lexemes for different grammar forms, examples of usage, etymology, etc.

7.5 The Sub-Entry

The sub-entry will have the same structure as the main dictionary entry.

7.6 The External Link

The external link will be used to connect the dictionary entry with external sources. For example, if a satirical newspaper is mentioned as the source, we will link it to its digitized form in the *Digital Repository of the Scientific Library of the Dubrovnik Libraries* (<https://zdur.dkd.hr/?pr=l&mr%5B27848%5D=a&ps=100>).

8. Conclusion

The primary objective of the Dictionary of the Dubrovnik Idiom project is to develop a diachronic, corpus-based, born-digital dictionary that encompasses the Dubrovnik idiom from the 16th to the 20th century. This paper outlines the methodology used to compile the corpus and the planned structure of the dictionary. It also describes the microstructure of the dictionary entries and discusses the anticipated challenges in creating the dictionary, such as the limitations of the available corpus and the task of reconstructing the vernacular from written sources. Additionally, it examines both the diachronic dimension of the dictionary and the dialectal aspects, particularly focusing on the accentuation and sound recordings of headwords.

For a project like this in Croatian philology, there is a significant need because a dictionary of the Dubrovnik idiom, created based on rigorous lexicographic principles, does not yet exist. Given the diachronic perspective, diversity, and

comprehensiveness of the texts that will comprise the corpus for this dictionary, as well as the pivotal role the Dubrovnik idiom has played in the standardization of the Croatian language and its importance in dialectological research, the corpus and the resulting e-dictionary will hold considerable scientific, expert, heritage, and cultural value. This value extends primarily to philologists and linguists but also to the broader Croatian public and an important cultural sector: the theatrical world, where actors often face challenges in conveying Dubrovnik texts and understanding the nuances of Dubrovnik sentences.

A special value will be brought by the sound recordings of the headwords. This element is a welcomed addition in contemporary lexicography, and in this project, it will present a significant challenge. Determining the correct accentuation will rely on systematic methods, linguistic and dialectological expertise, experience in working with the Dubrovnik idiom, and the authenticity provided by the project leader, who is a native speaker of the Dubrovnik idiom. This inclusion of sound recordings will enhance the dictionary's usability and authenticity, offering invaluable support to users in understanding and pronouncing the headwords accurately.

The geopolitical situation in the Dubrovnik area has led to demographic changes that are not favorable to the Dubrovnik idiom, so it is now heard very rarely in Dubrovnik. Therefore, this is the last opportunity to rescue its vocabulary from oblivion.

References

- ARJ – *Rječnik hrvatskoga ili srpskoga jezika* [Academic Dictionary of the Croatian or Serbian language] (1880–1976) I – XXIII. JAZU.
- Bencistà, A. (2021). *Vocabolario del vernacolo fiorentino toscano (con gli esempi delle voci da Dante ai moderni)*. Sarnus.
- Bojanić, M., & Trivunac, R. (2002). *Rječnik dubrovačkog govora*. Srpski dijalektološki zbornik XLIX. SANU.
- Brozović, D. (2005). Dva moguća fonološka tumačenja prozodijskoga sustava u cavtatskome govoru. Translation from English: H. Heffer. *Croatica et Slavica Iadertina*, 1(1), 19–30.
- Budmani, P. (1883). Dubrovački dijalekat, kako se sada govori. *Rad JAZU*, 65, 154–179.
- Despot, K. Š., & Möhrs, C. (2015). Pogled u e-leksikografiju. *Rasprave: Časopis Instituta za hrvatski jezik i jezikoslovlje*, 41(2), 329–353.
- Foji od Pazara (1901 – 1941). *Digital Repository of the Scientific Library of the Dubrovnik Libraries*. Retrieved April 3, 2024, from <https://zdur.dkd.hr/?pr=l&mr%5B27848%5D=a&ps=100>
- Gabrić-Bagarić, D., Horvat, M., Lovrić Jović, I., & Perić Gavrančić, S. (2012). *Ĵakov Mikalja. Blago jezika slovinskoga (1649./1651.): transkripcija i leksikografska interpretacija*. Institut za hrvatski jezik i jezikoslovlje.

- Hudeček, L., & Mihaljević, M. (2020a). The Croatian Web Dictionary – Mrežnik project – goals and achievements. *Rasprave*, 46(2), 645–667. <https://doi.org/10.31724/rihjj.46.2.11>
- Hudeček, L., & Mihaljević, M. (2020b). Collocations in the Croatian Web Dictionary – Mrežnik. *Slovenščina* 2.0., 8(2), 78–111. <https://doi.org/10.4312/slo2.0.2020.2.78-111>
- Lovrić Jović, I. (2006). Fonološka adaptacija talijanizama u dubrovačkim oporukama iz 17. i 18. stoljeća. *Rasprave Instituta za hrvatski jezik i jezikoslovlje*, 32(1), 173–192.
- Lovrić Jović, I. (2014). *O starome dubrovačkom govoru nazbilj. Jezična analiza dubrovačkih frančezarija*. Institut za hrvatski jezik i jezikoslovlje.
- Lovrić Jović, I. (2015). *Ja, Krsto Lučin Dubrovčanin, činim sad ovi testamenat. Jezična analiza dubrovačkih oporuka iz 17. i 18. stoljeća s transkripcijom i aneksnim rječnikom*. Institut za hrvatski jezik i jezikoslovlje.
- Lovrić Jović, I., & Jozić, Ž. (2016). *Libro od svetoga Vlaha. Jezična studija s rječnikom*. Institut za hrvatski jezik i jezikoslovlje.
- Metzeltin, M. (1992). La Dalmazia e l'Istria. In F. Bruni (Ed.), *L'italiano nelle regioni: lingua nazionale e identità regionali* (pp. 316–335). UTET.
- Morkovkin, V. V. (1996). The Theoretical Basis of a Systematic Approach to Dictionary Making. In: M. Gellerstam, J. Järborg, S.-G. Malmgren, K. Norén, L. Rogström, C. Rödger Pappmehl. *Euralex '96 Proceedings I–II*. Retrieved March 15, 2024, from https://www.euralex.org/elx_proceedings/Euralex1996_2/005_Valery%20V.%20Morkovkin%20-The%20Theoretical%20Basis%20of%20a%20Systematic%20Approach%20to.pdf
- Muljačić, Ž. (1962). Dalmatski elementi u mletački pisanim dubrovačkim dokumentima 14. stoljeća. Prilog raguzejskoj dijakronoj fonologiji i dalmatsko-mletačkoj konvergenciji. *Rad JAZU*, 327, 237–380.
- Muljačić, Ž. (1973). Problemi della simbiosi slavo-romanza nell'Adriatico. In M. Cortelazzo (Ed.), *Miscellanea II* (pp. 21–39). Facoltà di lingue e letterature straniere dell'Università di Trieste.
- Ničetić, J. (2012). *Dubrovački komini*. Self edition.
- Rešetar, M. (1933). Jezik Marina Držića. *Rad JAZU*, 248, 99–240.
- Rešetar, M. (1936). Jezik pjesama Ranjinina zbornika. *Rad JAZU*, 255, 77–220.
- Rešetar, M. (1941). Glavne osobine Gundulićeva jezika. *Rad JAZU*, 272, 1–44.
- Sočanac, L. (2004). *Hrvatsko-talijanski jezični dodiri (s rječnikom talijanizama u standardnome hrvatskom jeziku i dubrovačkoj dramskoj književnosti)*. Nakladni zavod Globus.
- Vulić, S. (2011). Štokavski hrvatski književni jezik. In A. Bičanić, R. Katičić, & J. Lisac (Eds.), *Povijest hrvatskoga jezika, 2. knjiga: 16. stoljeće*. (pp. 125–187). Društvo za promicanje hrvatske kulture i znanosti CROATICA.

Acknowledgments

This article was produced within the framework of the project “The Dictionary of the Dubrovnik Idiom”, funded by the European Union – NextGenerationEU. The views and opinions expressed are solely those of the authors, and these views do not necessarily reflect the official positions of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them.

Contact Information

Ivana Lovrić Jović

Institute for the Croatian Language

ilovric@ihjj.hr

Martina Kramarić

Institute for the Croatian Language

mkramar@ihjj.hr

Magdalena Majdak

DEFINING MEANINGS IN HISTORICAL DICTIONARIES

The Case of the Electronic Dictionary of the 17th- and 18th-Century Polish

Abstract This article addresses a crucial topic in lexicography and metalexicography, namely the challenge of defining meanings in historical dictionaries. Its aim is to present an overview of the criteria for crafting effective definitions by semanticists, logicians, and lexicographers, primarily tailored to meet the demands of contemporary dictionaries, to apply these principles to an academic historical dictionary of the Polish language and to assess the feasibility and justification of their implementation. The analysis encompasses: a) types of dictionary definitions; b) fundamental tenets of effective definitions, such as adequacy, substitutability, translatability, and analyticity; c) common pitfalls in the formulation of definitions (inadequate definitions – overly broad or narrow definitions, direct or indirect circular definitions, *ignotum per ignotius*); and d) the lexicographer's perspective, including the linguistic versus encyclopedic nature of definitions, the quandary of categorization (taxonomy) and valuation. The basis for the analysis is the *Electronic Dictionary of the 17th- and 18th-Century Polish*. The theory of definitions and its practical implementation in this dictionary are discussed. References are also made to other Polish historical-philological lexicons to elucidate comparable challenges and facilitate generalization.

Keywords lexicography; historical metalexicography; definition; historical dictionary; Polish

1. Historical Dictionaries and Their Characteristics

Historical dictionaries constitute a lexicographical endeavor that often exist on the periphery of contemporary metalexicographic discourse, likely for a very simple reason. Decisions to embark upon such an extensive project are infrequent, and the labor involved is both protracted and arduous. Historical dictionaries, extensive and paramount in their purview, typically evolve over decades. While bearing witness to emerging scholarly paradigms, they tend to assimilate these trends belatedly and with diminished impact. The pre-established conceptual framework obliges successive contributors to consistently adhere to the original guidelines to ensure the project's completion within the initially envisioned scope. While digital technologies have facilitated certain, sometimes automated, revisions in electronic dictionaries, the principal nature of lexicons cannot be subjected to fundamental changes.

The identification of specific characteristics and defining elements of historical dictionaries is significant, as it informs their structure and helps establish expectations for their content. These features may not always be immediately apparent due to their diversity and heterogeneity (Bielińska, 2020, pp. 318–321). Given the existence of numerous publications

titled ‘Historical Dictionary of [specific subject]’, such as *Historical Dictionary of Baroque Art and Architecture*, *Historical Dictionary of the Republic of Cameroon*, or *Historical Dictionary of Polish Cinema*, which *de facto* serve as lexicons documenting non-linguistic phenomena, it is imperative to clarify that this analysis focuses on a dictionary whose primary task is to describe linguistic facts.¹ Tentatively categorized as ‘historical-philological’, it distinguishes itself from historic dictionaries compiled in previous centuries.²

Based on criteria envisioned by various typologies (increasingly multifaceted and integrated, nowadays also because of digital technologies),³ a dictionary of this nature can be classified as scholarly, academic, general, monolingual, explanatory, documentary, or intended for professionals, notably qualified linguists, philologists or historians. Unlike commercial publications,⁴ it remains largely unaffected by publisher pressures (particularly when developed in electronic format), market dynamics,⁵ and is typically established under the auspices of the Academy of Sciences. What constitutes an essential aspect of metalexigraphic discourse is the user. Dictionaries are designed with users’ needs in mind, and significant emphasis is placed on ensuring their alignment with users’ capabilities.⁶ Nevertheless, as the authors of a well-known textbook observe, “This doesn’t imply a superficial concern with ‘user-friendliness’, but arises from our conviction that the content and design of every aspect of a dictionary must, centrally, take account of who the users will be and what they will use the dictionary for” (Atkins & Rundell, 2008, p. 5). In the context of a scholarly dictionary, the question remains whether its task is to be faithful to the theory or to the reader.

In discussing the purpose of scholarly dictionaries, Źmigrodzki (2003, p. 231) considers the “documentation of the linguistic material covered” and the concentration on “recording linguistic facts” their “primary function.” While historical dictionaries compiled, for example, in the 19th and 20th centuries prioritised the acquisition of material, which then had to be recorded and described, contemporary advancements in terms of access to sources through text corpora and digital libraries could potentially shift focus away from the processes of material collection and recording, which can now

¹ While Zgusta (1971, p. 199) argues that “It must also be remembered that the division of dictionaries into encyclopedic and linguistic ones is not necessarily an either-or matter,” Grochowski (2004, p. 9) presents a different approach.

² Although their definitions seem to differ slightly from those adopted in this article, Hartmann & James (2002, p. 68) also make a distinction between a ‘historical dictionary’ and a ‘historic dictionary’.

³ Aware of their complexity, Podhajecka (2009) offers a synthesis of many of these criteria.

⁴ “For all but scholarly or historical dictionaries, market forces come into play here: the new work will have to sell against existing dictionaries produced by competitor publishers” (Atkins & Rundell, 2008, p. 18).

⁵ “(...) there are two main types. The first is the scholarly and historical dictionary, a work often with few length constraints, and sometimes little pressure to complete within a specific time period – but also with a tendency to run out of money around letter C, or take 50 years to get there” (Atkins & Rundell, 2008, p. 31). It must be acknowledged that, unfortunately, there is a significant amount of truth in this somewhat blunt summary.

⁶ Given certain similarities to bilingual dictionaries, we could draw on the experiences of specialised and pedagogical dictionaries, whose authors distinguish various categories of users, such as laymen, semi-experts and experts (Bergenholtz & Tarp, 1995, p. 19). However, it is important to note that no contemporary editor can ever be a full expert in historical Polish.

be partially automated, to those of material description and semantic interpretation that could be further developed methodologically due to the advancement of theory. As Zgusta (1971, p. 201) says: “In a historical dictionary, the semantic developments are at least as important as those of the form of the word.” Given that historical-philological dictionaries are scholarly dictionary prototypes, it is expected that their definitions should adhere to academic standards. The feasibility of this expectation is explored below.

2. Defining Meanings in Historical-Philological Dictionaries

The formulation of definitions is considered one of the most difficult and also most creative stages of lexicographers’ work. There are many types of definitions, derived from diverse methodologies and perspectives on the world and language.⁷ Polish general lexicography has been dominated by the model proposed by Doroszewski (1958–1969) in his *Dictionary of Polish*. He delineated the ‘real-meaning’ definition, the structural-semantic definition, the scope definition, the synonymic definition, and the grammatical definition (Doroszewski, 1968). This model, along with the subsequent commentaries and discussions (e.g., Bartmiński, 1988; Bogusławski, 1988; Grochowski, 1988; Bańko, 2001; Piotrowski, 2001), has served as a reference point for all dictionaries, except for the COBUILD-inspired *Inny słownik języka polskiego* (‘A Different Dictionary of Polish’) edited by Mirosław Bańko, which features context-induced definitions.

Semanticists, logicians, and lexicographers specify the principles of effective (classical) definitions, such as adequacy, substitutability, translatability, and analyticity (employing semantically simpler elements in the *definiens* in relation to the *definiendum*). Additionally, they address fundamental errors in definitions, including definition inadequacy manifested by overly broad or narrow definitions, direct or indirect circular definitions, and the fallacy of explaining the unknown by the unknown (*ignotum per ignotius*). These guidelines are collected and discussed, *inter alia*, by Żmigrodzki, who as the editor of *Wielki Słownik Języka Polskiego Polskiej Akademii Nauk* (‘Polish Academy of Sciences Great Dictionary of Polish’; hereinafter: WSJP PAN), directly refers to the possible implementation of the semantic findings in the modern general dictionary, trying to apply them in practice.⁸

⁷ For example, see Svensen (2009, pp. 205–252).

⁸ In the *Principles for the preparation of WSJP PAN*, Żmigrodzki (2022, p. 30) says: “One of the recurring criticisms directed at the authors of subsequent dictionaries of Polish is the inconsistency of their definitions with specific semantic description proposals for respective linguistic units outlined in scholarly publications dedicated to lexical semantics. When embarking on the development of WSJP PAN, our aim was to address this issue and leverage the advancements of modern semantics to a greater extent than previously attempted. However, it must be acknowledged that this endeavour was not entirely successful. There are two primary reasons for this: firstly, the rather hermetic metalanguage of the definitions, rendering their straightforward integration into WSJP PAN impossible (the dictionary is intended for a broad audience); and secondly, the relatively limited availability and scope, compared to the planned number of entries in WSJP PAN, of units featuring descriptions (mainly abstract vocabulary and so-called functional expressions). Consequently, even if it were possible to adapt the scholarly explications to meet the requirements of the dictionary, a substantial portion of entries would necessitate the addition of such descriptions, or alternative methods of definition would have to be employed.” For more information on attempts to implement semantic theories into dictionaries, see also Żmigrodzki (2003, pp. 34–41) and Żmigrodzki (2008).

Svensen (2009, p. 205) states that “The purpose of the monolingual dictionary is to describe and explain the meaning and use of lexical items in one language by means of that language itself. The meaning of a certain object-language item is thus specified by means of one or more items belonging to the same language.” It is important to note that old language dictionaries encompass lexis from the ‘different world’, a bygone era. They often include terms that reflect concepts or contexts unfamiliar to contemporary users, which is why they are sometimes referred to as ‘peculiarly bilingual’ (Gruszczyński et al., 2018, p. 231). When preparing contemporary language dictionaries for native or non-native speakers, lexicographers have access to ample material and can rely on their own linguistic competence and knowledge to identify and describe the definitional and connotative features of entry words. However, when working on old language dictionaries, challenges emerge due to the limited availability of source material and the lack of the linguistic expertise required to provide comprehensive definitions with absolute certainty. Given a limited number of extant source texts, lexicographers are unable to create diagnostic contexts or conduct contradiction tests (Gruszczyński et al., 2018, p. 230). Moreover, crafting context-induced definitions poses significant risks due to the need for the partial reconstruction of meanings. Authors of definitions may rely on their scholarly intuition; however, a complete cultural and linguistic immersion is unattainable. While a linguistic competence can be developed to a certain level, its accuracy remains unverifiable. This raises questions regarding the practical implementation of methodological principles when compiling a historical dictionary. Which of them can be applied and to what extent?

In addition to typical theoretical and practical challenges faced by lexicographers, a shift is observed (at various levels, e.g., at the chronological one) in the descriptive layer with regards to the material described. It has significant implications, theoretically rendering the implementation of certain conditions for a good definition impossible. For instance, the 21st-century text layer of a contemporary definition would be incomprehensible and thus fraught with multiple logical errors to 17th-century language users. Moreover, achieving a semantic balance between the *definiendum* and the *definiens* to ensure that both are equivalent and thus substitutable, sometimes requires adding a pragmatic comment to the *definiens* (for example, the dictionary title specifies the chronological scope and thus acts as the label ‘17th and 18th century’). These observations lead to the conclusion that, logically, the *ignotum per ignotius* argument necessitates a distinct embrace of a modern perspective and a precise delineation with regards to what something is unknown or semantically simpler. Contemporary researchers lack full access to the semantic structure of old language. Compliance with the analyticity requirement poses a challenge, as it is difficult to determine whether contemporary lexical units are semantically simpler than their old counterparts.

Another issue that warrants attention is whether the definition of the entry word *czerwony* (‘red’) in *Słownik polszczyzny XVI wieku* (‘Dictionary of the 16th-century Polish Language’, hereinafter: SPXVI), described as ‘*mający barwę czerwoną lub zbliżoną do czerwonej*’ (‘having a color that is red or close to red’), should be regarded as an *idem per idem* error or a circular definition, or, similarly to the *ignotum per ignotius*, this rule should perhaps also be suspended due to the said temporal (and

stylistic) shift in the descriptive layer relative to what is described. Such changes may be observed in various aspects, such as categorisation (the addition of a hypernym that may not have existed or may have differed at the time, as evidenced by the shifts in the meaning of the word *sprzęt* ('equipment')), taxonomy (terms classified as 'a fish' in the 17th century could be reclassified as 'a reptile' in the 20th century or perhaps as something else entirely in the 21st century), and attribution of value/valuation (e.g., what and for whom constituted heresy). Our knowledge of the world is not the same as that of our predecessors; old meanings and connotations do not resonate the way their modern counterparts do today. There is a risk of anachronism when historical material is described from the perspective of a modern person whose perception of the world is influenced by contemporary scholarly categories and therefore differs from that of 17th- and 18th-century Polish language users.

This leads us to the question, frequently raised not only with regards to old language dictionaries, whether it is necessary to include definitions of words such as *matka* ('mother'), *krew* ('blood') or *mleko* ('milk'). Will contemporary language users seek definitions of these terms, for example, in SPXVI, or will they be more interested in information on how their meanings have changed compared to contemporary usage? The absence of a definition and the inclusion of context references only, sometimes seen in dictionaries, could indicate that the meaning of a given entry word overlaps with how it is understood today. Perhaps definitions of such terms could be 'outsourced' from a modern academic dictionary provided that it could be done systematically and upon authors' consent. Yet, this solution might not fully satisfy the editors of the definitions. A viable alternative could be to generate a database of dictionaries (and dictionary definitions) that would provide all definitions of a selected (standardised) entry word from numerous philological dictionaries, from Old Polish to Modern Polish, and allow to display them simultaneously for semantic comparison purposes. Currently, work is underway on the first database of this type in Poland – perhaps *Baza Historycznych Leksykonów Polskich BazHiLek* ('BazHiLek Database of Historical Polish Lexicons') could serve as a starting point for similar experiments.⁹

A review of the introductions or dictionary instructions of Polish historical-philological dictionaries – *Słownik staropolski* ('Old Polish Dictionary'), SPXVI and *Słownik języka polskiego XVII i 1. połowy XVIII wieku* ('Dictionary of the 17th-century and First Half of the 18th-century Polish'; hereinafter: SXVII) – conducted to assess their authors' declared adherence to semantic developments and their incorporation into the dictionary entries, reveals a primary focus on practical aspects. However, one must bear in mind that the concepts of these dictionaries were developed several decades ago, some as early as in the 19th century, making expectations of contemporary theory references not always justified. An example of an attempt to incorporate them into a historical-philological dictionary is *Elektroniczny Słownik Języka Polskiego XVII i XVIII wieku* ('Electronic Dictionary of the 17th- and 18th-century Polish').

⁹ The coordinating institution of the project is the Institute of the Polish Language of the Polish Academy of Sciences.

3. The Electronic Dictionary of the 17th- and 18th-Century Polish: Specific Problems and Solutions

The *Electronic Dictionary of the 17th- and 18th-Century Polish* (hereinafter: e-SXVII) offers a relatively robust theoretical framework for semantic description. This finding is unsurprising, given the editor and the recent development of the dictionary instructions (hereinafter: *Instructions*). Consequently, the editors of the *Instructions* (Gruszczyński & Adamiec, 2022) not only aimed to incorporate theoretical principles into the fundamentals of lexicographic descriptions but also had an opportunity to align with contemporary advancements in semantics. In Chapter 9, *Semantic Information*, the editors address key issues related to defining entries in the dictionary.¹⁰ They discuss the language, nature and types of definitions, the issue of authors' worldviews and specific problems regarding entry word descriptions by parts of speech.

Regarding the language of definitions, the editors argue that definitions should be formulated in contemporary Polish. They caution against the use of words not present in today's language to elucidate the meanings of lexical archaisms. This is why, for example, *angulowy* ('angular') is defined as '*tworzący kąt*' ('angle forming') rather than '*tworzący anguś*' ('angulus forming')¹¹. Authors of entries should prepare their definitions based on source contexts in which forms of the characterised lexical unit appear, while also drawing on their knowledge of history (culture and politics), historical-linguistic expertise and information from other historical-philological dictionaries (such as *Old Polish Dictionary* and SPXVI) (Section 9.3 of the *Instructions*).

There are two primary types of definitions: semantic and functional (Section 9.4). The most basic and frequently utilised form of semantic (objective) definitions is an analytic definition, wherein the *definiens* is a phrase composed of elements simpler than the *definiendum*. Based on this principle, it should be possible to substitute the unit constituting the *definiendum* with a sequence that is its *definiens*, without compromising the semantic correctness of the text (Section 9.4.1). Definitions of this type start with a hypernym followed by features distinguishing a given class. Sometimes, only a hypernym can be indicated when information about the meaning of the term is insufficient. Alternatively, a synonymic 'definition' may be employed, particularly for lexical archaisms where it is sufficient to provide the modern equivalent or equivalents like in *angul* ('angulus') – '1. mat. "kąt"' ('1. math. "angle"'). A combination of objective and synonymic definitions is also used. Semantic definitions must meet the translatability criterion.

¹⁰ Selected information is presented by quoting or paraphrasing excerpts from the *Instructions*.

¹¹ However, there are instances where 17th-century words not used in contemporary Polish become necessary, particularly when defining words derived from the names of designates that no longer exist (substantive archaisms), e.g., *arkabuzer* ('harquebusier'), *arkabuzerski* ('related to harquebus'), *arkabuźnik* ('harquebusier'). In such cases, definitions containing 17th-century words must follow a special formula to ensure that readers do not mistakenly assume that these words should be part of their (even passive) vocabulary (Section 9.2).

Scope definitions (Section 9.4.1.1.) indicate that the entry word pertains only to a limited range of referents, as exemplified by:

bąkać ('to hum'):

1. '*o owadach: brzęczeć, bzyczeć*' ('of insects: to buzz, to emit a low, continuous droning sound');
2. '*o człowieku: mówić cicho, szeptać*' ('of people: to utter an indistinct sound, to speak quietly, to hem');
3. '*o człowieku: wołać, krzyczeć; także z radości*' ('of people: to call out, to shout also out of joy');
4. '*o ssakach: wydawać donośny głos, ryczeć*' ('of mammals: to make a loud noise, to roar').

Functional definitions are metalinguistic in nature and do not adhere to the principle of translatability. They commonly refer to the grammatical features (grammatical definitions) or pragmatic features (pragmatic definitions) of the defined term, such as surnames, dog names or village names (Section 9.4.2).

The editors of e-SXVII emphasise their principle of avoiding encyclopaedic definitions. Definitions should focus on describing the properties of linguistic expressions rather than the characteristics of their referents. While elements of encyclopaedic knowledge are permissible and sometimes even necessary when defining historical terms (historisms), they can be included in the 'Additional Information' field (Section 9.4.1.2). The authors also address the issue of evaluation and worldviews in definitions, arguing that

Definitions of meanings should not be formulated from the point of view of the editor's system of values. In the case of e-SXVII, the most significant "risk" arises when defining terms related to religion. Definitions must not absolutise the concepts to which the defined lexical units refer. Instead, scope definitions such as "in Christianity" or "in Islam" should be used (Section 9.7).

The subsequent section of the *Instructions* (Section 9.8) delves into the specifics of defining various parts of speech. They outline specific formulas for defining nouns, such as agent nouns, proper nouns, feminine nouns, natural names, nouns describing nature and names of measures. The latter category is elaborated upon:

The following guidelines were accepted to standardise the definitions in entries pertaining to measures: the term "measure" (rather than "unit" or "measurement unit") is used in definitions with the information what it refers to – length, volume or mass (rather than "weight"); if it is possible or necessary, additional context is provided regarding the purpose of the measure, such as a volume measure for loose solids, a

volume measure for liquids or a mass measure used by pharmacists. Additionally, if possible, a conversion to modern measurement units should be offered; for example, see the definition of the term *funt* ('a pound') – '*miara masy o zróżnicowanej lokalnie wielkości równa 2 grzywnom, czyli 32 łutom; zwykle pomiędzy 0,4 a 0,5 kilograma*' ('a mass measure of locally varying magnitude, equal to 2 grzywnas or 32 lots; typically, between 0.4 and 0.5 kilogram') (Section 9.8.1).

In definitions of botanical terms,

we focus on the essential (particularly regarding the Baroque reality) characteristics of the described plant. We include information on its use if such knowledge is available based on source texts. Encyclopaedic knowledge and plant taxonomy descriptions are avoided. For example, [...] the definition of the term *dąb* ('oak') will be '*duże, rozłożyste drzewo liściaste o twardym drewnie, Quercus*' ('a large, spreading hardwood deciduous tree, *Quercus*') rather than '*Quercus Robur (Rost)*'; '*drzewo z rodziny bukowatych (Fagaceae)*' ('*Quercus Robur (Rost)*'; a tree in the beech family (*Fagaceae*')) (an encyclopaedic definition) or '*rodzaj drzewa liściastego*' ('a genus of deciduous tree') [...] (too general). We aim for a cognitive definition that also considers semantic connotations and non-scientific information which in our opinion could have been important to users of the 17th- and 18th-century Polish language. The contemporary Polish botanical name is provided at the end of the definition, after a comma (if it can be determined and is not identical with the entry word, as in the case of 'oak'), along with the Latin name (Section 9.8.4).

Importantly, the authors explicitly refer to cognitive definitions in terms of botanical names, while the names of primary colours are defined by means of prototypical objects associated with a specific color, e.g., *czzerwony* ('red') is defined as '*koloru krwi*' ('of the color of blood') (Section 9.9.4). This signals an expanded spectrum of definitions and methodologies in this historical dictionary compared to its predecessors. For example, see the excerpt on definitions in SXVII, whose section covering entry terms starting with letter 'A' was published before the dictionary was transformed into e-SXVII:

Various types of definitions are used depending on the requirements dictated by the material: explication definitions, e.g., *cyrkiel* ('compass') – '*przyrząd służący do kreślenia kół, do mierzenia powierzchni*' ('an instrument used for drawing circles and measuring surfaces'); structural and explication definitions in derivatives, e.g., *alembikowy* ('alembic') – '*dotyczący alembiku, będący częścią alembiku*' ('related to an alembic, a part of an alembic'); structural definitions in derivatives, typically in adjectival and adverbial entry words, and less frequently in noun entry words, e.g., *alabastrowy* ('alabaster') – '*przym. od "alabaster"*' ('adjective derived from [the noun] "alabaster"'); [...] scope definitions, e.g., *agitować, aitować* ('agitate') – '*o sprawach: poruszać, omawiać*' ('1. of matters: to move, to discuss'); translational definitions, e.g., *addycja* ('addition') – '*dodawanie*' ('summation') [...]. All types of definitions can be supplemented with synonyms, e.g., *absolucja* ('absolution') –

‘zwolnienie od kar kościelnych, rozgrzeszenie’ (‘exemption from church penalties, release, remission’). In some cases only synonyms are provided, e.g., *błyszczek* (‘trinket’) – ‘klejnot, błyskotka’ (‘an ornament, jewellery’) (SXVII, 1999–2004, p. XIV).

The dictionary analysed in this article, e-SXVII, might be the first Polish historical-philological lexicon whose *Instructions* refer to the theory of definition. While these instructions were developed based on practice, they also incorporate the knowledge of metalexicography. In view of certain insufficiencies in the structuralist paradigm highlighted by Gruszczyński, the dictionary opened to the inclusion of elements from other trends in semantics. A detailed discussion based on the material included in the dictionary may be found in his article (Gruszczyński et al., 2018).

4. Conclusions

Historical-philological dictionaries can be considered both fundamental and niche, but they are unique in many respects. According to Merkin (1983, p. 377), “The historical dictionary is therefore the most comprehensive type of scholarly academic dictionary, usually covering a national language with a long recorded history.” In their popular textbook on lexicography, *The Oxford Guide to Practical Lexicography*, Atkins & Rundell (2008, p. 121) highlight the specific character of these lexicons, by adding “except in the case of historical dictionaries” in most contexts relating to this type of works.

The fundamental question about the role of contemporary historical-philological dictionaries persists: Are they primarily intended to document and explain entry words, or also to help users decode their meanings (cf. Atkins & Rundell, 2008, p. 408), such as in the case of enthusiast of historical Polish writing novels or participating in reenactments? A plausible solution may entail crafting a smaller, popular dictionary, for example, a differential one that contains only meanings different from those used today, or an original dictionary of ‘false friends’ from the 17th and 18th centuries. However, it seems that what defines a dictionary as ‘academic’ is not hermetic definitions alone – especially since they are likely to be understood by an average educated user (e.g., e-SXVII complies with many ‘user-friendly’ requirements).

Views on the presence, role and shape of definitions in historical dictionaries are varied. For example, Atkins & Rundell (2008, p. 407) state: “It could be argued that definitions exist in order to catalogue the meanings in a language, and this is perhaps their chief function in a serious historical dictionary.” Conversely, Aitken (1971, p. 3) and Merkin (1983, p. 377), quoting the former author, believe that “the definitions and descriptive notes, which are also a normal feature of such dictionaries, may be regarded as fulfilling a somewhat secondary purpose, that of signposts or labels to the particular subset of quotations which follows” (similarly Żmigrodzki, 2003, p. 231). This shows that definitions may serve either a primary or secondary role.

Additionally, a concept articulated regarding another historical dictionary – *A Dictionary of the Older Scottish Tongue* (DOST) – warrants consideration:

It is the quotations and not the definitions which carry the main burden of what the dictionary has to offer and are its primary concern. In effect, the dictionary is an alphabetically indexed and analysed collection of quotations, and in principle it should be possible for the user of the dictionary to work out for himself the word's semantic range, the chronological and regional distributions and so on, simply from the quotations and their arrangement without recourse to the definition at all' (Aitken, 1973, p. 6).

This raises questions not only about the role of definitions but also about that of lexicographers. Today, after years of employing text corpora for lexicographic purposes, it is well established that the approach of “all that is required is to feed a large historical corpus into a computer, press a button or click a mouse, and out comes the ready-made historical dictionary” (Pajz, 2000, p. 249) may not be sufficient. The mere accessibility of quotations does not equate to the reconstruction of meanings, and ordinary language users are often unable to deduce meanings despite having open access to extensive material.

The question arises whether the role of a definition in a historical dictionary should be reduced to merely serving as a label for a subset of quotations illustrating the same meaning, or whether it should resemble a classical scholarly definition or the reconstruction of what Apresjan refers to as ‘naïve worldview’. In our case, it would involve reconstructing the old worldview (with all due proportion), whereby the 17th- and 18th-century material would be approached like folk material and described from the perspective of language users from that era. These ideas are implemented in practice, for example, by *Słownik symboli i stereotypów ludowych* (‘Dictionary of Folk Stereotypes and Symbols’), where *wieloryb* (‘a whale’) is categorised as a fish (Bartmiński, 1988, p. 20). Perhaps a two-stage definition should be considered. Some entries could offer two definitions, allowing users to switch between the one written from the contemporary perspective and another reflecting the worldview typical of the Baroque or Enlightenment.

Durkin (2016, pp. 2–3) states that “the historical dictionary, rightly or wrongly regarded by many as the ‘prestige’ dictionary par excellence, is also associated by some with old-fashioned methods, academicism, stuffiness, or even a certain hauteur.” An opportunity for a revision of similar diagnoses may be found in resources like *Elektroniczny Korpus Tekstów Polskich z XVII i XVIII w.* (‘The Electronic Corpus of 17th- and 18th-century Polish Texts’), completed in December 2023 and currently numbering nearly 27 million tokens (Gruszczyński et al., 2023). This corpus, together with its digital index, has significantly increased the number of existing attestations of word units from that period. Its gradual integration with e-SXVII (Bronikowska et al., 2020; Bilińska-Brynk & Rodek, 2020) has presented a variety of challenges and opportunities for lexicographers to improve both theory and practice. Similarly, the BazHiLek database may also prove helpful in this regard. It may be necessary to

develop additional rules for creating definitions that will correspond to contemporary semantic developments while remaining functional, useful and economical from the perspective of lexicography. The search for alternative ways of defining meanings in academic dictionaries of historical language is bound to contribute to the evolution of general lexicographic principles.

Translated from Polish by Małgorzata Sobczak.

References

- Aitken, A. J. (1971). Historical dictionaries and the computer. In R. A. Wisbey (Ed.), *The Computer in Literary and Linguistic Research*. University of Cambridge.
- Aitken, A. J. (1973, 2015). Sense-analysis for a historical dictionary. In H. Scholler, & J. Reidy (Eds.), *Lexicography and Dialect Geography* (pp. 5–16). Festgabe for Hans Kurath. Reproduced in Aitken, A. J. & Macafee, C. (Ed.). (2015). *Collected Writings on the Scots Language*. Scots Language Centre.
- Bańko, M. (2001). *Z pogranicza leksykografii i językoznawstwa. Studia o słowniku jednojęzycznym*. Wydział Polonistyki Uniwersytetu Warszawskiego.
- Bartmiński, J. (1988). Definicja kognitywna jako narzędzie opisu konotacji słowa. In J. Bartmiński (Ed.), *Konotacja* (pp. 169–183). Uniwersytet Marii-Curie Skłodowskiej.
- Bartmiński, J. (1988). Słownik ludowych stereotypów językowych. Założenia ogólne. In *Etnolingwistyka*, 1, 11–34.
- Bergenholtz, H., & Tarp, S. (Eds.). (1995). *Manual of Specialised Lexicography. The Preparation of Specialised Dictionaries*. John Benjamins.
- Bilińska-Brynk, J., & Rodek, E. (2020). Paper Quotation Slips to the *Electronic Dictionary of the 17th- and 18th-Century Polish* – Digital Index and its Integration with the Dictionary. In Z. Gavriilidou, M. Mitsiaki, & A. Fliatouras (Eds.), *Proceedings of the XIX EURALEX Congress: Lexicography for Inclusion* (pp. 465–470). Democritus University of Thrace.
- Bogusławski, A. (1988). *Język w słowniku. Desiderata semantyczne do wielkiego słownika polszczyzny*. Zakład Narodowy im. Ossolińskich.
- Bronikowska, R., Majdak, M., Wieczorek, A., & Żółtak, M. (2020). The *Electronic Dictionary of the 17th- and 18th-century Polish* – Towards the Open Formula Asset of the Historical Vocabulary. In Z. Gavriilidou, M. Mitsiaki, & A. Fliatouras (Eds.), *Proceedings of the XIX EURALEX Congress: Lexicography for Inclusion* (pp. 471–475). Democritus University of Thrace.
- Durkin, P. (Ed.). (2016). *The Oxford Handbook of Lexicography*. Oxford University Press.
- Doroszewski, W. (1968). O definiowaniu znaczeń wyrazów. *Poradnik Językowy*, 5, 236–246.
- Grochowski, M. (1988). Podstawowe zasady definiowania wyrazów w słowniku jednojęzycznym. In W. Lubaś (Ed.), *Wokół słownika współczesnego języka polskiego* (pp. 45–62). Zakład Narodowy im. Ossolińskich.

Grochowski, M. (2004). Informacja semantyczna w słownikach popularnych i naukowych. Postulaty metodologiczne. *Poradnik Językowy*, 9, 8–18.

Gruszczyński, W., Adamiec, D., Majdak, M., & Morawski, W. (2018). Wybrane problemy semantyki historycznej. Definiowanie znaczeń w słowniku dawnej polszczyzny. In M. Pastuch, & M. Siuciak (Eds.) in cooperation with K. Wąsińskiej & W. Wilczek, *Historia języka w XXI wieku. Stan i perspektywy* (pp. 229–242). Wydawnictwo Uniwersytetu Śląskiego.

Gruszczyński, W., & Adamiec, D. (2022). *Instrukcja redakcyjna dla twórców Elektronicznego słownika języka polskiego XVII i XVIII wieku. Materiały wewnętrzne*.

Gruszczyński, W., Adamiec, D., Bronikowska, R., Kieraś, W., Modrzejewski, E., Wieczorek, A., & Woliński, M. (2023). The Electronic Corpus of 17th- and 18th-century Polish Texts. *Language Resources and Evaluation*, 56, 309–332.

Merkin, R. (1983). Historical Dictionaries and the Computer – Another View. In S. Allén, P. Corbin, R. R. K. Hartmann, F. J. Hausmann, H.-P. Kromann, O. Reichmann, & L. Zgusta (Eds.), *Proceedings of the 1st EURALEX International Congress* (pp. 377–384). Max Niemeyer Verlag Tübingen.

Pajz, J. (2000). Making Historical Dictionaries by Computer. In U. Heid, S. Evert, E. Lehmann, & Ch. Rohrer (Eds.), *Proceedings of the Ninth EURALEX International Congress, EURALEX 2000* (pp. 249–259). Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

Piotrowski, T. (2001). *Zrozumieć leksykografię*. Wydawnictwo Naukowe PWN.

Podhajecka, M. (2009). Historical Development of Lexicographical Genres. Some Methodological Issues. In R. W. McConchie, A. Honkapohja, & J. Tyrkkö (Eds.), *Selected Proceedings of the 2008 Symposium on New Approaches in English Historical Lexis (HEL-LEX 2)* (pp. 153–170). Somerville.

Svensén, B. (2009). *A Handbook of Lexicography*. Cambridge University Press.

Zgusta, L. (1971). *Manual of Lexicography*. Mouton.

Żmigrodzki, P. (2003). *Wprowadzenie do leksykografii polskiej*. Wydawnictwo Uniwersytetu Śląskiego.

Żmigrodzki, P. (2008). *Słowo – słownik – rzeczywistość. Z problemów leksykografii i metaleksykografii*. Wydawnictwo Lexis.

Żmigrodzki, P. (2023). *Wielki słownik języka polskiego PAN. Zasady opracowania*. Retrieved May 30, 2024, from <https://wsjp.pl/page/zasady>

Dictionaries:

Bańko, M. (Ed.). (2000). *Inny słownik języka polskiego*. Wydawnictwo Naukowe PWN.

Bartmiński, J., & Niebrzegowska-Bartminska, S. (Eds.). (1996–2004). *Słownik stereotypów i symboli ludowych*. Uniwersytet Marii Curie-Skłodowskiej. Instytut Filologii Polskiej.

Bielińska, M. (Ed.). (2020). *Leksykografia. Słownik specjalistyczny*. TAIWPN Universitas.

Doroszewski, W. (Ed.). (1958–1969). *Słownik języka polskiego PAN*. Wydawnictwo Naukowe PWN.

Gruszczyński, W. (Ed.). (2004–). *Elektroniczny słownik języka polskiego XVII i XVIII wieku*. Instytut Języka Polskiego PAN. <https://sxvii.pl>

Hartman, R. R. K., & James, G. (2002). *Dictionary of Lexicography*. Routledge.

Mrowcewicz, K., & Potoniec P. (Eds.). (1956–). *Słownik polszczyzny XVI wieku*. Instytut Badań Literackich.

Siekierska, K. (Ed.). (1996–2004). *Słownik języka polskiego XVII i 1. połowy XVIII wieku*. Instytut Języka Polskiego PAN.

Urbańczyk, S. (Ed.). (1953–2002). *Słownik staropolski*. Instytut Języka Polskiego PAN.

Żmigrodzki, P. (2007–). *Wielki słownik języka polskiego PAN*. Instytut Języka Polskiego PAN. <https://wsjp.pl>

Acknowledgements

I would like to extend my sincere gratitude to Prof. Włodzimierz Gruszczyński for his feedback and insightful remarks regarding this paper.

The research has been supported by the National Programme for the Development of Humanities (NPRH) under the project ‘Third-level digitisation of large 17th- and 18th-century dictionaries: The development of the Database of Historical Polish Lexicons’ (no. NPRH/DN/SP/0003/2023/12) funded by the Ministry of Science and Higher Education.

Contact information

Magdalena Majdak

Institute of Polish Language, Polish Academy of Sciences
magdalena.majdak@ijp.pan.pl

DIALECT DICTIONARY AND LEXICALIZATION OF DIALECT PHENOMENA

Abstract This paper investigates Czech territorial dialect lexicography, particularly the incorporation of lexicalized phonological and morphological phenomena into differential dialect dictionaries. It examines the methodologies used in current Czech dialect dictionary creation. The study analyzes corpora of the Czech National Corpus (CNC), including ORAL v1, ORTOFON v2, and DIALEKT, to trace the lexicalization of dialect features. Observing lexicalization in Czech dialects through corpus analysis elucidates its various phases, from initial frequency increase to subsequent restriction in specific lexical units. Through illustrative examples (semi-semiconsonantal *u*, hard *l*, the nominative plural ending of the masculine animatum noun *-i*), the study sheds light on lexicalization's evolution and distribution in contemporary spoken Czech. Additionally, it addresses challenges in documenting regular and irregular dialectal variations and proposes that lexicalized forms, even when they do not alter the lemma, should be included in differential dictionaries. Such an approach would enhance the representation of dialectal differences, contributing to a more comprehensive understanding of Czech dialectology.

Keywords Czech; dialects; lexicalization; dialect dictionary

1. Dialect Lexicography

From the outset of dialectological research, studying and describing dialect vocabularies has been central. Dictionaries of territorial dialects vary in their approach, including territorial scope, time span, and entry organization, aiming for either comprehensive or differential vocabularies (Sochová, 1995).

Regarding Czech dialects, the following types exist (Vojtová, 2002): differential, alphabetically ordered dictionaries (Kellner; 1949, Lamprecht, 1963), onomasiological glossaries combined with differential dictionaries (Bachmannová, 1997; Holub, 2001; Sochová, 2001), complete subject explanatory dictionaries (Roudný, 1952). The ongoing Dictionary of Czech Dialects aims to map all Czech dialects as a relatively complete, non-differential dictionary (Ireinová, 2019). It is an undeniably valuable, but extremely time-consuming project (the author's team has been working on it since 2011), which for the time being publishes electronically the processed part of the passwords (Slovník nářečí českého jazyka, 2016–). Due to the complexity of mapping entire vocabularies, most local dialect dictionaries use a differential approach, either alone or with subject dictionaries. The differential dialect dictionary (Sochová, 1995, p. 260) captures lexical units that are specifically dialectal. These include namely dialect-specific lexical units (they do not have a standard counterpart), proper lexical dialectisms (they have a standard counterpart), semantic dialectisms (they have a different meaning than the standard form), frequency dialectisms (their frequency

is lower in standard language), stylistic dialectisms (the word is similar to the standard form, but its form is varied according to the phonological or morphological dialectal features) and contextual dialectisms (phrases). Challenges arise in including lexicalized phonetic or morphological phenomena in differential dictionaries. This study tracks lexicalization in spoken Czech corpora, focusing on integrating these phenomena into differential dialect dictionaries.

2. Corpora of Spoken Czech in the Czech National Corpus

The Czech National Corpus (CNC) showcases spoken Czech's regional diversity through several corpora. For observing lexicalization of lexical and morphological dialectal phenomena, the most relevant corpora are ORAL v1 (Kopřivová et al.), ORTOFON v2 (Kopřivová et al.), and the specialized corpus DIALEKT (Goláňová et al.). ORAL v1 (5.4 million tokens) captures contemporary informal spoken Czech from 2002–2011. Its single transcription plane highlights regular deviations from written pronunciation. ORTOFON v2 (2.1 million tokens) represents informal spoken Czech from 2012–2019, captured on two transcription levels (orthographic and phonetic) linked to audio recordings. Both corpora include adult speakers from all generations across the Czech Republic. DIALEKT (223,000 tokens) documents traditional territorial dialects, mostly through monologic speeches by the oldest generation of speakers (born between 1875 and 1957) from all Czech dialect areas, including Czech language islands in Poland. By comparing contemporary spoken Czech (ORAL v1 and ORTOFON v2) with traditional dialects (DIALEKT), we will investigate the presence of lexicalization in selected phenomena.

3. Development of Czech Dialects and Lexicalization of Phonetic and Morphological Phenomena

Czech dialects, like those of other languages, resulted from historical development. The main driver of divergent linguistic evolution was feudal fragmentation. However, from the 19th century onwards, population migration and the growth of large cities initiated dialect leveling, where narrowly regional features receded and broader regional features persisted, leading to dialect convergence. Differentiating phonetic and morphological phenomena do not vanish immediately across the entire vocabulary but undergo a gradual process. Regular and productive phenomena often undergo lexicalization, remaining limited to specific lexical units or word forms for a time. Eventually, they may either disappear entirely or persist as linguistic relics.

3.1 Examples of Lexicalization Observed in Spoken CNC Corpora

3.1.1 Semi-Semiconsonantal *ʉ*

The pronunciation of 'v' as a semi-semiconsonantal *ʉ* was originally common in the dialect regions of northeastern and central Bohemia (Český jazykový atlas 5, 2005, p. 429). The DIALEKT corpus documents 143 lemmas where *ʉ* occurs in

Northeast Bohemian and Central Bohemian dialects. Among these lemmas, the following are the most frequent (with an occurrence rate of over 50 instances per million words):

1. povídat	brzy	všechn	zrovna	opravdu ¹
tell	soon	every	just	really

Most frequent word forms (i.p.m. over 50):

2. pouďá	pouďám	dřiu	zrouna ůsechno	vopraudu ²
povídá	povídám	dřív	zrovna všechno	pravdu
				(standard forms)
he tells	I tell	earlier just	everything	really

The ORAL and ORTOFON corpora identically record *u* (in the subcorpus of speakers coming from the northeastern and central Bohemia region) already lexicalized in the following forms (i.p.m. over 1), ORTOFON:

3. (v)opraudu (i.p.m. 24,64)	prauda (8,96)	zrouna (6,72)
opravdu	pravda	zrovna (standard forms)
really	true	just

ORAL:

4. (v)opraudu (i.p.m. 4,69)	prauda (4,35)	houno (3,35)
opravdu	pravda	hovno (standard forms)
really	true	shit

The occurrence of *u* pronunciation is not limited to the oldest generation as the bearers of the historical form of the dialect, but is distributed evenly in both general corpora among all generations of speakers (from 18 to 72 years old). In this case, the DIALEKT corpus captures the initial process of lexicalization and the ORAL and ORTOFON corpora represent the advanced stage.

3.1.2 Hard *t*

The historical pronunciation of the hard *t*, primarily characteristic of eastern Moravia and Silesia, was recorded in most of the peripheral dialect areas during the 20th century (Český jazykový atlas, 2005, p. 172). Examples of its lexicalization can be found in the Northeast Bohemian dialect area in the form of pronunciation relics, documented in the DIALEKT corpus (with an occurrence rate of 1,022.1 instances per million words). Among the 48 lemmas, *být* – “to be” is significantly more prevalent (with an occurrence rate of 58.03 instances per million), while others have rates below

¹ povídat (i.p.m. 306,25), brzy (122,5), všechn (96,71), zrovna (70,92), opravdu (61,25)

² pouďá (i.p.m. 119,28), pouďám (74,15), dřiu (70,92), zrouna (70,92), ůsechno (54,8), vopraudu (51,58)

10, indicating random occurrences. Specifically, verbs in the past participle³, where *l* serves as the formant, exhibit this phenomenon:

5.	byl	šel	dal
	byl	šel	dal (standard forms)
	he was	he went	he gave

Occurrences of other forms are mostly random and have rates below 10 i.p.m. The lexicalization of this phenomenon is evident. While the ORTOFON corpus documents lexicalized pronunciation in the forms of the verb *být* (“to be”) in other regions, it is no longer observed in Northeastern Bohemia. Additionally, the ORAL corpus does not record the pronunciation of the hard *l*.

3.1.3 The Nominative Plural Ending of the Masculine Animatum Noun -í

Among the morphological phenomena, lexicalization is evident in the nominative plural ending of the masculine animatum noun -í, which originally characterized the dialect in the southwestern half of Bohemia as a regular and typical feature (Český jazykový atlas 4, 2002, p. 152). The DIALEKT corpus demonstrates its consistent occurrence across the vocabulary (with 61 lemmas). Higher frequencies (over 10 instances per million words) are observed in the following forms:

6.	klucí	chlapcí	Řemců ⁴	Američanů	vojáci ⁵
	kluci	chlapci	Řemci	Američani	vojáci (standard forms)
	boys	boys	Germans	Americans	soldiers

In the ORAL and ORTOFON corpora, the phenomenon is already clearly lexicalized in the forms of *klucí* (“boys”, with occurrences of 14.78 per million words in ORAL and 8.59 per million words in ORTOFON). However, in other instances, the phenomenon appears randomly (with occurrences of less than 1 per million words), often exclusively in the speech of a single speaker, demonstrating the influence of idiolect.

3.2 Detecting Lexicalization Phases Through Corpus Analysis

These examples show how observing the distribution of phenomena in the corpora of older dialect forms and contemporary spoken Czech can help us reveal lexicalization and its phases. The first phases are typically characterized by an increase in the frequency of the observed phenomenon for certain lexical units or their groups. In the next phase, the phenomenon is restricted exclusively to certain lexical units, the frequency of which may gradually decrease.

³ tag VpYS---XR-AA– (i.p.m. 119,28), VpNS---XR-AA– (22,57), VpQW---XR-AA– (12,9), VpYS---XR-NA– (12,9)

⁴ The high frequency of the words *Řemci*, *Američanů* and *vojáci* is influenced by the choice of topics the dialect speakers monologue about, in this case the topic of World War II.

⁵ *klucí* (i.p.m. 106,38), *chlapci* (54,8), *Řemci* (45,13), *Američanů* (29,01), *vojáci* (12,9)

4. Capturing Lexicalized Phenomena in a Differential Dictionary

How are lexicalized phonological and morphological dialect phenomena incorporated into differential dialect dictionaries? The current practice for creating Czech dialect dictionaries involves comparing entries with *Slovník spisovného jazyka českého* (Dictionary of the written Czech language, SSJČ) (Bachmannová, 2016, p. 13). Entries are included if they are absent from the SSJČ, have spelling differences, or exhibit vowel quantity variations. Additionally, words labeled as dialectal, regional, or folk in the SSJČ are also included. In other dictionaries, only significant lexical differences are noted, excluding minor variations like vowel quantity (Bachmannová, 1998, p. 11).

Typically, words that differ from the standard form only by a regular variation in the basic form are usually not included in dictionaries. Such cases are usually discussed in introductory essays that describe the dialectal features being mapped. Words with regular dialectal changes in non-lemma forms are often covered in the morphological descriptions of the dialect rather than as dictionary entries.

More challenging are lexicalized (irregular or residual) phonological or morphological changes. These are generally not included as separate entries but are mentioned in introductory sections, such as (Bachmannová, 2016, p. 9): Archaic Feature: Pronunciation of the hard “l” (l̥) in past participles (*byl, přišel, měla⁶*) and occasionally elsewhere (*ke stolu, na dluh, hořt, vokoľ⁷*).

If a lexicalized phenomenon occurs in a word’s lemma, it should be a separate dictionary entry. For instance (Bachmannová, 2016, p. 286):

7. povdat (pouđat)
povídat (standard form)
to tell

However, if the phenomenon appears only in non-lemma forms, such lexical units are not included. Previously regular but now lexicalized forms, such as *klucí* (3.1.3), do not appear in dictionaries because their lemmata match the standard form and the change occurs only in other forms.

These cases should be included in differential dialect dictionaries, even if they do not show lemma differences. Lexicalized differential forms in other word forms should be noted and commented on within the entry. The question remains whether to list them under the non-differential lemma:

8. kluk – lexikalizovaný tvar nom. pl. *klucí*
kluk – lexicalized form nom. pl. *klucí*

⁶ He was, he came, she had.

⁷ To the table, on credit, simply, around.

or introduce the entry with the differential form:

9. klucí – nom. pl. lexikalizovaný tvar (zákl. tvar *kluk*)
klucí – nom. pl. lexicalized form (lemma *kluk*)

The first approach aligns with existing entries, while the second allows easier retrieval.

5. Conclusion

Tracking the occurrence of phenomena in various spoken language corpora reveals the progressive lexicalization of originally regular phonological or morphological phenomena. Including these lexicalized forms in differential dialect dictionaries, especially when they appear in non-lemma forms, would provide a more comprehensive dialect description. Even if the lemma matches the standard form, describing the dialectal difference within the entry enhances the overall depiction of the dialect.

References

- Bachmannová, J. (1997). *Podkrkonošský slovník*. Academia.
- Bachmannová, J. (2016). *Slovník podkrkonošského nářečí*. Academia.
- Český jazykový atlas 4 (2002). Academia.
- Český jazykový atlas 5 (2005). Academia.
- Golánová, H., Waclawičová, M., & Lukeš, D. (2021). DIALEKT: nářeční korpus, verze 2 z 23. 12. 2021. Ústav Českého národního korpusu FF UK. Retrieved January 10, 2024, from <http://www.korpus.cz>
- Holub, Z. (2001). *Lexikon nejjižnějšího úseku českých nářečí*. Aleš Čeněk.
- Ireinová, M. (2019). O Slovníku nářečí českého jazyka. In É. Császári, & M. Imrichová (Eds), *Király Péter 100. Tanulmánykötet Király Péter tiszteletére I.* (pp. 223–229). Szilv Filológiai Tanszék.
- Kellner, A. (1949). *Východolašská nářečí II*. Dialektologická komise při Matici moravské.
- Kopřivová, M., Laubeová, Z., Lukeš, D., Poukarová, P., & Škarpová, M. (2020). ORTOFON v2: Korpus neformální mluvené češtiny s víceúrovňovým přepisem. Ústav Českého národního korpusu FF UK. Retrieved January 10, 2024, from <http://www.korpus.cz>
- Kopřivová, M., Lukeš, D., Komrsková, Z., Poukarová, P., Waclawičová, M., Benešová, L., & Křen, M. (2017). ORAL: korpus neformální mluvené češtiny, verze 1 z 2. 6. 2017. Ústav Českého národního korpusu FF UK. Retrieved January 10, 2024, from <http://www.korpus.cz>
- Lamprecht, A. (1963). *Slovník středoopavského nářečí*. Krajské nakladatelství Ostrava.

Roudný, J. (1952). *Slovník nářečí podle zápisů z Úlibic na Jičínsku*. [Candidate dissertation, Univerzita Karlova v Praze].

Slovník nářečí českého jazyka (2016–). Dialektologické oddělení Ústavu pro jazyk český AV ČR, v. v. i. Retrieved January 10, 2024, from <http://sncj.ujc.cas.cz>

Slovník spisovného jazyka českého (1960–1971). Academia.

Sochová, Z. (1995). Nářeční lexikografie. In F. Čermák, & R. Blatná (Eds.), *Manuál lexikografie* (pp. 249–264). H&H.

Sochová, Z. (2001). *Lašská slovní zásoba*. Academia.

Vojtová, J. (2002). Perspektivy nářeční lexikografie. *Sborník prací Filozofické fakulty brněnské univerzity. A, Řada jazykovědná*, 51(A50), 105–110. <https://hdl.handle.net/11222.digilib/100913>

Acknowledgements

This work has been supported by Charles University Research Centre program No. 24/SSH/009.

Contact information

Martina Waclawicová

Institute of the Czech National Corpus, Faculty of Arts, Charles University in Prague
martina.waclawicova@ff.cuni.cz

Leonardo Zilio and Besim Kabashi

USING NEURAL MACHINE TRANSLATION FOR NORMALISING HISTORICAL DOCUMENTS

Abstract The work with historical documents presents many challenges, not only because some sources are not well preserved, but also because grammar and spelling rules from older times were not always consistent. Still, these texts remain as a rich source of information from our history, and we could greatly benefit from the information that can be extracted from them. At the same time, the lack of spelling and grammatical consistency poses a problem for the application of computational tools, so most of the analysis work is done manually. To overcome this lack of consistency, researchers started normalising the spelling of historical documents, as this increases the performance of modern tools. Spelling normalisation is, however, also carried out manually most of the time. In this paper, we present some experiments that were done for automatically normalising historical documents in two languages: Portuguese and Albanian. Leveraging state-of-the-art large language models that were pre-trained for translation, we used corpora that were carefully curated and manually normalised to train new computational models. These models can automatically normalise documents in these languages, achieving new state-of-the-art BLEU scores above 90 for Portuguese, and up to 59 for Albanian, beating the task baselines.

Keywords historical linguistics; Albanian; Portuguese; spelling normalisation; computational linguistics; large language models; neural machine translation

1. Introduction

Historical documents have proven time and time again to be an enormous challenge for the automatic processing and extraction of information (cf. Quaresma & Finatto, 2020; Vieira et al., 2021; Cameron et al., 2022; Zilio et al., 2022; Zilio et al., 2023; Vikomir & Herndon, 2024). While there is progress in the field of natural language processing (NLP) for historical texts, most of the tools are still developed with a focus on modern iterations of language, and only a few studies have been dedicated to the computational processing of historical documents in their original spellings.

To help mitigate issues caused by spelling differences, researchers started resorting to normalising the writing of historical documents (Piotrowski, 2012; Bollmann & Søgaard, 2016; Bawden et al., 2022; Vieira et al., 2024); that is, they started updating the spelling of historical texts using modern-day orthographic rules. This enables modern tools to extract data from historical documents with higher precision, as word forms in modern spelling have higher probability of having been present in their training data. However, normalisation work still is mainly done manually and is very time consuming.

Having this scenario as background, the objective of this study is to explore more automatised ways of producing a normalised text from a source historical document.

As such, this paper extends the work of several researchers that already explored machine translation methods for text normalisation (cf. Bollmann & Søgaard, 2016; Domingo & Casacuberta Nolla, 2018; Domingo & Casacuberta Nolla, 2019), with the novelty that we investigate more recently introduced transformer-based large language models (LLMs). Another novelty is that we concentrate our efforts on languages that have been relatively less studied in terms of historical documents and automatic normalisation: Portuguese and Albanian. Our method is based on fine-tuning already existing neural machine translation (NMT) models, so we do not need large amounts of data, and a corpus composed of a few thousand words is already enough to achieve good results.

The main contributions of this paper are the following:

- An experiment using fine-tuned NMT models for the task of spelling normalisation of historical documents.
- The release of a new parallel corpus of Albanian texts, containing 12,677 tokens of original spelling and 12,836 tokens of normalised spelling, which can be used for fine-tuning NMT models.
- A comparison of performance for automatic normalisation in Portuguese and Albanian.

The remainder of the paper is organised in the following way: Section 2 discusses previous work done in the area of normalisation and automatic normalisation, while also commenting on previous efforts of applying natural language processing tools to historical documents; Section 3 presents the corpora that we used as basis for this study; Section 4 describes the methodology and the LLMs that we used for the automatic normalisation of historical documents; on Section 5 we present the results of the experiments for both languages, discussing the performance of individual LLMs; finally, on Section 6 we present our final remarks, discussing the main achievements of this study.

2. Related Work

Many authors have faced the challenge of using NLP methods for working with historical documents. Although we have not found any register for works dealing with historical data in Albanian in this setting, some studies dealt with the automatic spelling normalisation using machine translation models as basis. However, most of these studies do not take advantage of the more recent transformer models, and none that we could find uses the fine-tuning of large language models as a means to achieve spelling normalisation.

More recent studies involving NMT for spelling normalisation usually rely on an encoder-decoder, character-based architecture based on long-short term memory (LSTM) models (cf. Bollmann & Søgaard, 2016; Domingo & Casacuberta Nolla, 2018; Domingo & Casacuberta Nolla, 2019). While these studies make sense, by modelling

the spelling normalisation problem as a character-based replacement, Tang et al. (2018) have already hinted that subword tokens can provide a better solution to character-based models. This points to the use of subword embeddings as basis for the automatic normalisation. However, the study by Tang et al. (2018) is the only one that we could find that tests subwords for this task, and neither Portuguese nor Albanian are among the tested languages.

In the subject of languages in focus, studies on automatic normalisation are concentrated around European languages. Bollmann (2019) developed a large comparison of automatic spelling normalisation methods for English, German, Hungarian, Icelandic, Portuguese, Slovene, Spanish, and Swedish. Other studies focused on fewer languages, such as the work of Domingo & Casacuberta (2019) for Slovene and Spanish, Bawden et al. (2022) for French, and Robertson (2017) for English, German, Icelandic, and Swedish. For Portuguese, we are only aware of the above-mentioned work by Bollmann (2019), who used a corpus of letters from the 15th to 19th century that was made available by the Post Scriptum project (CLUL, 2014). More recently, researchers at the University of Évora started working with text normalisation. Cameron et al. (2023) developed a categorisation of variants, which can support the normalisation of historical Portuguese texts, and Olival et al. (2023) described the normalisation of six documents belonging to the *Parish Memories* collection. In addition to text normalisation, there are studies that use historical data in different NLP tasks, such as named-entity recognition (cf. Ehrmann et al., 2023) and textual complexity (cf. Zilio et al., 2023).

As for the Albanian language we could not find any register of work being done on any automatised task related to historical documents, and this study is possibly the first of its kind, presenting a new normalised dataset and results for automatic spelling normalisation of historical documents.

3. Historical Corpora

3.1 The Portuguese Corpus

The Portuguese sample was collected from three medical handbooks published in the 18th century. It was originally released by Zilio et al. (2024) and is freely available for download on Github in several formats, including train, validation and test splits as TSV files¹. Table 1 presents the number of tokens, sentences and overall statistics of the historical corpus as observed with AntConc (Anthony, 2005). As we can see, it is by no means a large corpus (especially considering today's age of big data), containing a little over 24 thousand tokens, but the results are nonetheless very impressive, as we will discuss later in Section 5.

This corpus was used mainly as a source of comparison for the work done on the Albanian language. This way, we add results from two new models that were not tested in the original paper (as can be seen in Section 5), and we discuss the results of a low-resourced language when compared to a higher-resourced language.

¹ Available at: https://github.com/uebelsetzer/automatic_normalisation_of_historical_documents.

Table 1: Description of the Portuguese historical corpus, considering both the original spelling and the normalised spelling. The type ratio represents the ratio between types in the original corpus and the types in the normalised corpus and is an indicator of spelling variation in the original documents. T/S = average tokens per sentence

	Portuguese historical corpus	
	Original	Normalised
Tokens	24,504	24,815
Types	5,584	5,341
Type ratio	N/A	1.05
Sentences	453	453
T/S	54.09	54.78

3.2 The Albanian Corpus

The Albanian corpus consisted of samples selected from three religious books from the 17th century. All three samples were written by Pjetër Budi (1566–1622), whose Italian name variant was Pietro Budi. He was born in Gur i Bardhë, in Mat, a region in central Albania and he was an Albanian Catholic bishop.

The three selected books were published for the first time between 1618 and 1621. The first sample (ca. 20 pages) was selected from *Doctrina Cristiana*, or *Doktrina e Kërshtenë* (*The Christian Doctrine*). It was originally published in Rome, 1618 and has 286 pages (cf. Svane, 1985). The second sample (ca. 20 pages) was extracted from *Rituale Romanum*, or *Rituali Roman* (*Roman Ritual*). This book was published in Rome in 1621 and has 376 pages (cf. Svane, 1986a). The third sample (ca. 20 pages) was selected from *Speculum Confessionis*, or *Pasëqyra e t'rrëfyemit* (*The Mirror of Confession*). It was published in Rome in 1621 and has 409 pages (cf. Svane, 1986b). As usual at that time, the texts were printed following the spelling used by the author, which in some cases is not consistent.

Table 2: Description of the Albanian historical corpus, considering both the original spelling and the normalised spelling. The type ratio represents the ratio between types in the original corpus and the types in the normalised corpus. T/S = average tokens per sentence

	Albanian historical corpus	
	Original	Normalised
Tokens	12,677	12,836
Types	2,216	2,235
Type ratio	N/A	0.99
Sentences	307	307
T/S	41.29	41.81

Table 2 displays quantitative information about the whole corpus in its original and normalised versions. As it can be seen, the normalised corpus contains more types than the original, which indicates that, perhaps because of Albanian being a language

with casus declination, the normalised version ends up having different spellings for words that were spelled in the same way in the historical documents. Another explanation for this could be that many words had to be split into separate words in the normalised version, resulting in more types.

4. Methodology

The methodology that we used in this study is very straightforward. We started by manually normalising the spelling of historical documents and then aligned the original and the normalised versions of the same text. These alignments are mostly done at the sentence level, however, for Albanian, we preferred to split sentences also at semi-colons, to have more segments available for training. These aligned segments were then used for fine-tuning existing transformer-based neural machine translation models, which were then applied to automatically normalise historical documents in the trained languages and domains.

The process of normalising a historical document consists of updating the spelling of words in a way that they match modern spelling standards of the language in focus. For Portuguese, the documents were normalised using modern Brazilian Portuguese spelling standards (cf. Zilio et al., 2024), while, for Albanian, the modern spelling norm (1972) was used. Here are examples of original and normalised versions of two short segments in Portuguese (Examples 1 and 2) and in Albanian (Examples 3 and 4):

1. He opinião praticada pelo Cirurgiaõ Mór do nosso Hospital Real de Elvas Francisco Xavier , cuja sciencia he notoria . **(Original)**
2. É opinião praticada pelo Cirurgiã-Mor do nosso Hospital Real de Elvas Francisco Xavier, cuja ciência é notória. **(Normalised)**
3. E kshtu mb atë ças vojta , zuna fill përserii ; **(Original)**
4. E kështu më atë çast vajta , zura fill përsëri ; **(Normalised)**

As it can be seen from these examples, there are some different choices that were made by the normalisers of both languages. For instance, in Portuguese, the punctuation was also normalised to correspond to the standards of modern Brazilian Portuguese, while in Albanian the punctuation was kept as in the original text.

The two corpora presented in Section 3 were split into train, development, and test sets for each of the languages. This is standard procedure for preparing the corpus for fine-tuning NMT models. The two corpora were split a bit differently, based on the composition of each corpus.

In the case of Portuguese, full texts (i.e., individual chapters) were preserved in the train and development sets, and in the test set. This resulted in a division, based on number of tokens, of approximately 74% for training, 8% for development, and 18% for testing, as can be seen in Table 3.

Table 3: Train, development and test sets derived from the Portuguese historical corpus, considering both the original spelling and the normalised spelling. T/S = average tokens per sentence

	Train		Development		Test	
	Original	Normalised	Original	Normalised	Original	Normalised
Tokens	18,047	18,286	2,038	2,067	4,419	4,462
Types	3,386	3,213	826	803	1,372	1,325
Type Ratio	-	1.05	-	1.03	-	1.04
Sentences	342	342	38	38	73	73
T/S	52.77	53.47	53.63	54.39	60.53	61.12

Table 4 shows the data splits for Albanian. Because there is no clear subdivision in the texts that were used (except for coming from the three different sources), we simply split segments from each of the three samples into train, development and test set, making sure that segments from sentences that were split at semi-colon remained on the same set. This resulted in a more even split of 70% of the segments for training (~74% of tokens), 10% for development (~6% of tokens), and 20% for testing (~21% of tokens).

Table 4: Train, development and test sets derived from the Albanian historical corpus, considering both the original spelling and the normalised spelling. T/Seg = average tokens per segment

	Train		Development		Test	
	Original	Normalised	Original	Normalised	Original	Normalised
Tokens	9,345	9,484	709	695	2,623	2657
Types	1,830	1,838	299	291	777	770
Type Ratio	N/A	1	N/A	1.03	N/A	1.01
Segments	241	241	35	35	70	70
T/Seg	38.77	39.35	20.26	19.86	37.47	37.96

4.1 NMT models

The models we used in this study had to contain both Portuguese and Albanian in their set of trained languages, and they also needed to be trained in a many-to-many fashion, i.e., both languages had to have been presented both as source and as target during the training process. This severely limited our scope, as there are not many models that include Albanian in their training data, so we had to settle for three models, which were directly downloaded from the Huggingface repository (<https://huggingface.co/>):

- **m2m 100 (m2m100)²**: resulting from the work of Fan et al. (2021), the m2m models include a hundred languages trained in a many-to-many fashion, i.e., each language can be translated directly into the other, without using intermediate languages as point of reference. The base model contains 1.2

² For more information: https://huggingface.co/facebook/m2m100_418M.

billion parameters, but it was not supported by our graphics card, so we used a smaller version, with 418 million parameters.

- **NLLB-200 (NLLB)**³: the *No language left behind* paper (NLLB Team, 2022) had a huge impact on the machine translation community, as it offers the largest combination of languages to date, while focusing on low-resourced ones. While it contains many languages in its training data, it is also a less focused model, and while it works in advancing the machine translation state of the art for some low-resourced languages, it might not perform as brilliantly for highly resourced ones, such as Portuguese, when compared to models that include fewer languages. Again, we could not use the model's base version, because our graphics card did not support it, we instead used a distilled version that contains 600 million parameters.
- **SMaLL 100 (sm100)**⁴: as the name suggests, this is a much smaller version of the m2m100 model. Developed by Mohammadshahi et al. (2022), this model still contains all 100 languages trained in a many-to-many fashion, but its distillation process was focused on preserving information for low-resourced languages. Again, we have to observe how this focused distillation might have an impact on its performance for both Portuguese and Albanian.

All models were trained using an Nvidia RTX 4090 with 24GB of RAM using the standard *transformers* library as provided by Huggingface⁵ (Wolf et al. 2020) for Python. All models were fine-tuned with the same parameters: learning rate of 2e-5, weight decay of 0.01, and 100 epochs (but only the best epoch was saved at the end based on BLEU score performance on the development set). Batch sizes had to be changed according to the model size, so that the GPU could load them into memory: 8 for m2m and sm100, and 6 for NLLB.

4.2 Evaluation metric

We evaluated all models using the BLEU score metric (Papineni et al., 2002), which is commonly used for assessing the performance of machine translation engines. There is much criticism against BLEU scores, because the metric indeed has several shortcomings when used for evaluating machine translation. BLEU is a metric that compares the number of n-grams in the target text with reference text(s) and produces a score from 0 to 100.

One of the downsides of BLEU is reproducibility. Because any size and combination of n-grams can be used, its reproducibility depends on a well-described methodology. To address this issue, Post (2018) introduced SacreBLEU, which is a standard way of obtaining BLEU scores. In this paper, we used the *evaluate* library's implementation of SacreBLEU for assessing BLEU scores for each model.

³ For more information: <https://huggingface.co/facebook/nllb-200-distilled-600M>.

⁴ For more information: <https://huggingface.co/alirezamsh/small100>.

⁵ The *transformers* library is available on Github: <https://github.com/huggingface/transformers>.

Another downside of BLEU is that it assesses the target text using one or multiple reference texts. These references themselves may be questionable target texts, and they do not necessarily invalidate other equally correct translation options for a given source text. As such, a lower BLEU score might be just a reflex of different translation choices in the references. In our case, this issue is mitigated, because most of the time there is no alternative correct option for a given token in the normalisation pipeline. Most words in historical documents can only be normalised to one single form using modern spelling standards.

As a baseline for comparing the improvement that was achieved by the NMT models for the normalisation task, we used the original text and scored it against the normalised version. This way we have a baseline for what is the BLEU score in case no change is made in the original document. As for assessing the internal improvement of the NMT models with fine-tuning, we used their respective non-fine-tuned versions as baseline.

5. Results

Table 5 presents the results in terms of BLEU scores for off-the-shelf models, which serve as baselines for the fine-tuning, and fine-tuned models. The task baseline, i.e., the BLEU score of the original test set when compared to the normalised test set, is shown in the middle, as no off-the-shelf model was able to score higher than this baseline.

For Portuguese, the experiment with fine-tuning m2m100 provided a new state-of-the-art for automatic spelling normalisation in historical medical documents in the proposed dataset, as it beat the previous best score achieved by a fine-tuned mBART model. The fine-tuned m2m100 model not only had the best score, but its non-fine-tuned version also achieved the best off-the-shelf score. Models that were more focused on low-resourced languages, such as NLLB and sm100 achieved inferior results in this test set.

For Albanian, the results show that sm100 was the best model, technically tied with m2m100 because of rounding, but actually 0.0061 BLEU points ahead. As we can see from Table 5, normalising Albanian historical documents from the 17th-century is a much more complex task than 18th-century Portuguese, as shown by the task baseline of 16.78 BLEU points, compared to 62.07 in Portuguese. The off-the-shelf models had a really hard time achieving any sort of readable output, and sm100 had the best score, with 6 BLEU points, which goes to show how far behind low-resourced languages still are when compared to higher resourced ones. Upon fine-tuning, we can see a good improvement, moving the scores up to 59 BLEU points. This improvement for m2m100 and sm100 is actually higher for Albanian than for Portuguese, as it represents a jump of more than 50 BLEU points, while in Portuguese the jump was a bit over 30 BLEU points.

Table 5: BLEU-score results on Portuguese and Albanian test sets, separated into off-the-shelf models, task baseline, and fine-tuned models. * Scores for mBART were reproduced from Zilio et al. (2024) because this model had the former best performance for normalisation in the Portuguese dataset; unfortunately, this model does not cover Albanian.

Model	SacreBLEU	
	18 th -century Portuguese	17 th -century Albanian
Off-the-shelf models		
mBART*	30.73	N/A
m2m100	57.98	3.15
NLLB	40.64	2.97
sm100	57.44	6.00
Task baseline		
No changes in test set	62.07	16.78
Fine-tuned models		
mBART*	88.20	N/A
m2m100	90.31	59.08
NLLB	83.65	43.39
sm100	89.04	59.08

6. Discussion and Final Remarks

In this paper we leveraged large language models (LLMs) that were pre-trained for neural machine translation (NMT) to explore the feasibility of automatising the spelling normalisation of historical documents. As basis for our experiments, we used an existing corpus of medical texts in Portuguese and developed a completely new corpus of Albanian religious texts. This new Albanian dataset is being released with this paper and is available on Github⁶.

We selected three LLMs that were pre-trained for NMT and included both Portuguese and Albanian as source and target languages. Each of these models includes a variety of languages, and we purposefully selected two models that were variants of each other, where one of the variants was a distilled version focused on low-resourced languages, so we could see the influence of removing embeddings for certain tokens when dealing with a higher resourced language, such as Portuguese, when compared to a low-resourced language, such as Albanian.

For Portuguese, a new state of the art in spelling normalisation was achieved by using the m2m100 model, which surpassed the performance of mBART, the previous state of the art for medical documents (cf. Zilio et al., 2024) by more than 2 BLEU-score points. This model also achieved the best off-the-shelf result (i.e., without fine-tuning), even if it still did not score above the task baseline. The distillation

⁶ Available at: <https://github.com/uebelsetzer/NMT4AlbanianNormalisation>.

performed on m2m100 to create sm100 indeed reduced the size of the generated models, making them lighter to use: m2m100 models occupy 5.41GB of space, while sm100 models have a size of 3.72GB (approximately 69% of the size). However, this came at a small cost of performance for Portuguese, as the result, although marginally better than mBART, was still more than one BLEU-score point below its larger version.

For Albanian, the models achieved more than 40 BLEU points above the task baseline after fine-tuning. The distillation process focused on low-resourced languages did not seem to have much impact on the final scores, as sm100 did not achieve a much higher score than its non-distilled counterpart, and results from off-the-shelf m2m100 with 1.2 billion parameters (5.47 BLEU points) indicate that it could surpass sm100 upon fine-tuning.

As far as limitations go, there are a few elements that should be mentioned. One of the main restrictions of this paper is that we had a single RTX 4090 for running the experiments, which did not allow us to fine-tune larger models, such as the m2m100 1.2GB⁷ or the NLLB 200 3.3GB⁸. These models, when tested on the Portuguese dataset using their off-the-shelf versions, produced BLEU scores of 63.35 and 9.18, respectively. As such, this larger m2m100 was the first off-the-shelf model to score above the baseline, which points to potential state-of-the-art performance after fine-tuning, and, although the large NLLB model had an abysmal score before fine-tuning, it is a model that has great potential for improvement, as it has information collected from a couple hundred languages.

Another limitation was the size of the datasets. Because all the spelling normalisation work had to be done manually, collecting more data is a very time-consuming work. As such, the Albanian corpus, which started being constructed much later than the Portuguese corpus, is not as large, and thus there is probably room for improvement if we can manage to have a larger corpus with samples from more authors.

Even with these limitations, this study managed to contribute with a new dataset and new state-of-the-art models for automatic spelling normalisation for both Portuguese and Albanian. We hope that these resources will help advance the studies in Historical Linguistics, and we expect to use them to extend the existing datasets, as these models can potentially speed up the normalisation process, by turning it into post-editing instead of a fully manual task.

References

Anthony, L. (2005). AntConc: design and development of a freeware corpus analysis toolkit for the technical writing classroom. In *IPCC 2005. Proceedings. International Professional Communication Conference*, 2005. (pp. 729–737). IEEE.

⁷ For more information: https://huggingface.co/facebook/m2m100_1.2B.

⁸ For more information: <https://huggingface.co/facebook/nllb-200-3.3B>.

- Bawden, R., Poinhos, J., Kogkitsidou, E., Gambette, P., Sagot, B., & Gabay, S. (2022). Automatic normalisation of early Modern French. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 3354–3366).
- Bollmann, M. (2019). A Large-Scale Comparison of Historical Text Normalization Systems. In *2019 Conference of the North American Chapter of the Association for Computational Linguistics, Minneapolis, Minnesota, June 7, 2019*. (pp. 3885–3898). Association for Computational Linguistics.
- Bollmann, M., & Søgaard, A. (2016). Improving historical spelling normalization with bi-directional LSTMs and multi-task learning. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (pp. 131–139).
- Cameron, H. F., Olival, F., & Vieira, R. (2023). Planear a normalização automática: tipologia de variação gráfica do corpus das Memórias Paroquiais (1758). *LaborHistórico*, Rio de Janeiro, ISSN, 2359-6910.
- CLUL. (2014). P.S. Post Scriptum. *Arquivo Digital de Escrita Quotidiana em Portugal e Espanha na Época Moderna*. <https://www.clul.ulisboa.pt/projeto/ps-post-scriptum>.
- Domingo, M., & Casacuberta Nolla, F. (2018). Spelling normalization of historical documents by using a machine translation approach. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation: 28–30 May 2018, Universitat d’Alacant, Alacant, Spain* (pp. 129–137). European Association for Machine Translation.
- Domingo, M., & Casacuberta Nolla, F. (2019). Enriching Character-Based Neural Machine Translation with Modern Documents for Achieving an Orthography Consistency in Historical Documents. In *International Conference on Image Analysis and Processing* (pp. 59–69).
- Ehrmann, M., Hamdi, A., Pontes, E. L., Romanello, M., & Doucet, A. (2023). Named entity recognition and classification in historical documents: A survey. *ACM Computing Surveys*, 56(2), 1–47.
- Fan, A., Bhosale, S., Schwenk, H., Ma, Z., El-Kishky, A., Goyal, S., et al. (2021). Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107), 1–48. <https://doi.org/10.48550/arXiv.2010.11125>
- Mohammadshahi, A., Nikoulina, V., Bérard, A., Brun, C., Henderson, J., & Besacier, L. (2022). SmaLL-100: Introducing Shallow Multilingual Machine Translation Model for Low-Resource Languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 8348–8359).
- NLLB Team (2022). No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.
- Olival, F., Cameron, H. F., Farrica, F., & Vieira, R. (2023). As Memórias Paroquiais (1758) do atual concelho de Vila Viçosa. *Callipole: revista de Cultura*, 29, 85–128.
- Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (pp. 311–318).

Piotrowski, M. (2012). *Natural language processing for historical texts*. Morgan & Claypool Publishers.

Post, M. (2018). A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers* (pp. 186–191), Brussels, Belgium.

Quaresma, P., & Finatto, M. J. B. (2020). Information Extraction from Historical Texts: a Case Study. In *DHandNLP @ PROPOR* (pp. 49–56).

Robertson, A. (2021). *Automatic Normalisation of Historical Text*. PhD thesis, School of Informatics, University of Edinburgh.

Svane, Gunnar (Ed.). (1985). *Pjetër Budi. Dottrina christiana (1618). With a transcription into modern orthography and a concordance (Sprog og Mennesker 9)*, Institut for Lingvistik, Aarhus Universitet. Århus.

Svane, Gunnar (Ed.). (1986a). *Pjetër Budi. Rituale Romanum (1621). With a transcription into modern orthography and a concordance (Sprog og Mennesker 11)*, Institut for Lingvistik, Aarhus Universitet. Århus.

Svane, Gunnar (Ed.). (1986b). *Pjetër Budi. Speculum Confessionis (1621). With a transcription into modern orthography and a concordance (Sprog og Mennesker 13)*, Institut for Lingvistik, Aarhus Universitet. Århus.

Tang, G., Cap, F., Pettersson, E., & Nivre, J. (2018). An Evaluation of Neural Machine Translation Models on Historical Spelling Normalization. In *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 1320–1331).

Tang, Y., Tran, C., Li, X., Chen, P. J., Goyal, N., Chaudhary, V., et al. (2020). Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*. <https://doi.org/10.48550/arXiv.2008.00401>

Vieira, R., Olival, F., Cameron, H., Santos, J., Sequeira, O., & Santos, I. (2021). Enriching the 1758 Portuguese Parish Memories (Alentejo) with named entities. *Journal of Open Humanities Data*, 7(20).

Vieira, R., Cameron, H., Olival, F., Farrica, F., Finatto, M., Banza, A., Ribeiro, A., Trojahn, C. (2024) PLN e Humanidades Digitais. In H. M. Caseli, & M. G. V. Nunes (Org.). *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português. 2 ed. BPLN, 2024*. <https://brasileiraspln.com/livro-pln/2a-edicao/parte-dominios/cap-humanidades-digitais/cap-humanidades-digitais.html>

Vilkomir, K., & Herndon, N. (2024). Challenges of Automatic Document Processing with Historical Data. In *Proceedings of the 2024 ACM Southeast Conference on ZZZ* (pp. 50–59).

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., et al. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: system demonstrations* (pp. 38–45).

Zilio, L., & Finatto, M. J. B. (2022). Named Entity Recognition Applied to Portuguese Texts from the XVIII Century. In *Proceedings of the Second Workshop on Digital Humanities and*

Natural Language Processing (2nd DHandNLP 2022) co-located with International Conference on the Computational Processing of Portuguese (PROPOR 2022), Fortaleza, Brazil, 21st March, 2022 (pp. 1–10). CEUR.

Zilio, L., Finatto, M., Vieira, R., & Quaresma, P. (2023). A natural language processing approach to complexity assessment of 18th-century health literature. *Domínios de Linguagem*, 17.

Zilio, L., Lazzari, R. R., & Finatto, M. J. B. (2024) Can rules still beat neural networks? The case of automatic normalisation for 18th-century Portuguese texts. In *Proceedings of the International Conference on the Computational treatment of Portuguese*, volume 2 (pp. 86–95). <https://aclanthology.org/2024.propor-2.0.pdf>

Acknowledgements

We would like to thank the Chair of Computational Corpus Linguistic at the Friedrich-Alexander-Universität Erlangen-Nürnberg for the support.

Contact information

Leonardo Zilio

Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
leonardo.zilio@fau.de

Besim Kabashi

Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
besim.kabashi@fau.de

