Philipp Stöckle, Wolfgang Koppensteiner,
Katharina Korecky-Kröll, Maria Luisa Krapp, Tanja Wissik (eds.)

LEXICOGRAPHY IN THE AGE OF AI

BOOK OF ABSTRACTS
OF THE XXII EURALEX INTERNATIONAL CONGRESS 2026

29 September - 3 October 2026, Vienna | Austria

EURALEX
EURALEX AUSTRIA 2026
AUSTRIAN ACADEMY OF SCIENCES PRESS
ÖAW Austrian Academy of Sciences
austrian centre for digital humanities

Editors: Philipp Stöckle, Wolfgang Koppensteiner, Katharina Korecky-Kröll, Maria Luisa Krapp, Tanja Wissik

Cover & title page design: Sandra Lehecka

Some rights reserved.
Copyright © Austrian Academy of Sciences, Vienna 2026
Typesetting: Sunrise Setting, Brixham, UK
ISBN: 978-3-7001-5139-5
https://epub.oeaw.ac.at/5139-5
https://verlag.oeaw.ac.at

FOREWORD

This book contains the abstracts of keynotes, talks, posters, and software demonstrations presented at the XXII EURALEX International Congress, Vienna, Austria, 29 September – 3 October, 2026.

The abstracts were either submitted and accepted as extended abstracts or were taken as short abstracts from full papers accepted for publication in the Proceedings of the XXII EURALEX International Congress.

Part I contains an overview of all keynotes, extended abstracts and short abstracts of talks, posters, and software demonstrations, as well as the workshops. Each entry in these lists is linked to the corresponding abstract. Part II contains the abstracts of keynote presentations. Part III contains the extended abstracts, Part IV the short abstracts, both listed in alphabetical order by the authors' last names. The pre-conference workshop descriptions are included in Part V. Finally, an alphabetical index contains all authors' names.

The main topic of the XXII EURALEX International Congress is **Lexicography in the Age of AI**. The rapid development of artificial intelligence (AI) and natural language processing (NLP) has begun to transform the ways in which lexicographic data are compiled, analysed, and presented. Large language models, advanced corpus tools, and other machine learning methods offer unprecedented opportunities to accelerate tasks such as the extraction of lexical information, the induction of word senses, or the drafting of definitions. They also make possible new forms of user interaction, including conversational and personalized dictionaries that can adapt to different audiences or integrate multimodal content such as text, sound, and images.

At the same time, these technologies pose serious challenges that go beyond purely technical concerns. Automatically generated content raises questions of reliability, bias, and transparency, while the integration of opaque systems into lexicographic workflows risks undermining scholarly standards of quality and accountability. The ecological impact of large-scale AI models must also be taken into consideration, raising doubts about their long-term sustainability in resource-intensive projects such as lexicography. Finally, the introduction of AI into the field invites reflection on the role of the lexicographer: how can human expertise remain central when automation is increasingly pervasive, and what kinds of collaborations between humans and machines are both effective and responsible?

The conference provides a forum for discussing these promises and risks. We bring together professional lexicographers, publishers, researchers, software developers, and anyone interested in dictionaries to consider both the potential of AI for creating more dynamic, interactive dictionaries, and the ethical and ecological responsibilities that come with it. While many contributions engage directly with the overarching theme of AI and its implications for lexicography, the conference programme also reflects the full breadth of the field of lexicography, encompassing diverse theoretical, methodological, and applied perspectives beyond the conference's main focus.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

Sponsors xvi

Programme Committee xvii

Scientific Committee xvii

PART I OVERVIEW

Keynotes xxii

Extended Abstracts xxii

Short Abstracts xxvi

Workshops xxxi

PART II KEYNOTES

Robert Lew 2

Kate Wild 3

Andreas Baumann 4

Ana Salgado 5

Jan Hajič 6

PART III EXTENDED ABSTRACTS

Vassiliki Afentoulidou, Ioannis Tziogkas, Christos Papanagiotou, Aretousa Giannakou & Anastasia Christofidou 8
AI-ASSISTED LEXICOGRAPHY IN MODERN GREEK: Neology and Genre

Mikel Alonso 11
AI-ASSISTED INFORMATION EXTRACTION FROM LARRAMENDI'S TRILINGUAL DICTIONARY (1745)

Maria Ammari 14
PEDAGOGY AS THE CORE OF EDUCATION: A Domain-Analytic Route to a Corpus-based Dictionary

Michela Bandini, Silvia Piccini, Andrea Bellandi & Emiliano Giovannetti 17
TOWARD A NARRATIVE-ORIENTED DICTIONARY: Computational Modelling of Youth Language in Socially Vulnerable Contexts

Biljana Božinovski, Makimi Kano, Peter Gobel & Mihaela Franjić 20
LEXICOGRAPHIC FUNCTIONS WITHOUT DICTIONARIES: GenAI-Mediated Lexical Problem-Solving Among Non-English-Major Tertiary EFL Students

Gabriela Brůhová & Veronika Raušová 23
THE "HELL OF A" CONSTRUCTION IN CASUAL WRITTEN ENGLISH ONLINE: A Synchronic Functional Account

Edmond Cane 25
LEARNING LIKE HUMANS, PARSING UNLIKE MACHINES:
A Construction-Based Approach to Lexical Annotation

Elena Castaldi 27
INFLECTED-FORM SEARCH IN ANCIENT GREEK DICTIONARIES:
A Form-To-Lemma Access Model

Elina Chadjipapa, Zoe Gavriilidou, Maria Mitsiaki & Ifigeneia Dosi 29
CHATBOTS IN PEDAGOGICAL LEXICOGRAPHY: Evaluating AI-generated dictionary entries from three LLMs

John J. Chew III, Emma Ferrett & Anastasia Riehl 32
CANADIAN ENGLISH DICTIONARY: A Progress Report

Luisa Cimander & Petra Storjohann 34
TRACING NEOLOGISMS IN NEO-DISCOURSES: Diachronic and Cross-Domain Perspectives

Václav Cvrček, Martin Stluka & Klára Pivoňková 36
TRACKING EIGHT CENTURIES: The Czech Monitor Corpus as a Resource for Diachronic Lexical Research

Veronique De Tier, Katrien Depuydt, Koen Mertens & Mathieu Fannee 38
FROM PILOT TO PRACTICE: Two Years of Progress in Developing a Shared Infrastructure for Dutch Dialect Dictionaries

Kristina Š. Despot, Ana Ostroški Anić & Matej Klemen 42
BEYOND BINARY METAPHOR DETECTION: Dataset-Driven Evaluation of Conceptual Metaphor Type Recognition in Croatian and Slovene

Emma Ferrett, Amir Ghorvei & Anastasia Riehl 45
A DICTIONARY FOR ALL CANADIANS: Establishing an Inclusive Standard for the Canadian English Dictionary

Emmanuel Sylvain Fomat 47
CAMFRANGLAIS IN THE AGE OF ARTIFICIAL INTELLIGENCE: Lexicographic Challenges, Opportunities and a Human-AI Approach to Documenting Linguistic Hybridity in Multilingual Cameroon

Daniele Franceschi 51
GENDER REPRESENTATION AND EDITORIAL RESPONSIBILITY IN THE OED:
A case study of *female*, *feminine*, *male* and *masculine*

Anne-Kathrin Gärtig-Bressan & Pia Carmela Lombardi 53
BILINGUAL DICTIONARIES WITH ITALIAN IN THE 19TH AND 20TH CENTURIES: The ALON portal

Marius Glebus 56
CHOOSING OPTIMAL AI PROMPT INSTRUCTIONS FOR NEOLOGISM DETECTION

Wanjing Han & Gilles-Maurice de Schryver 58
CAN LLMS CAPTURE CHINA'S EMERGING ECONOMIC NEOCONCEPTS LEXICOGRAPHICALLY?

Wanjing Han & Gilles-Maurice de Schryver 61
CORPUS-BASED AND AI-ASSISTED DESIGN OF A CHINESE–ENGLISH BUSINESS DICTIONARY

Table of Contents

EXTENDED ABSTRACTS

Kris Heylen, Katrien Depuydt, Carole Tiberius & Jesse de Does 67
BUILDING A SENSE INVENTORY FOR THE CENTRAL LEXICOGRAPHIC INFRASTRUCTURE OF DUTCH

Kris Heylen, Ilan Kernerman, Judit Freixa, Tarja Heinonen, Hae-Yun Jung, Henrik Lorentzen, Kilim Nam, Benjamin Rouxel, Ana Salgado, Miguel Sánchez Ibáñez, Petra Storjohann & Lars Trap-Jensen 70
THE ENEOLI MATRIX OF NEOLOGISM SELECTION CRITERIA: A Comparative Framework for Lexicographic Practice

Louise Holmer 73
NEOLOGY MILLS? AI-Influenced Newspaper Articles and their Implications for Lexicology and Lexicography

Željko Jozić, Marijana Horvat & Martina Horvat 75
DIGITIZATION OF ADAM PATAČIĆ'S MANUSCRIPT DICTIONARY

Hae-Yun Jung, Jinsan An, Kilim Nam & Jun Lee 78
FROM NEOLOGISM COLLECTION TO LEXICOGRAPHIC DESCRIPTION: A Corpus-LLM-Lexicographer Triangulation Framework

Ilan Kernerman, Federica Vezzani, Anna Vacalopoulou, John Humbley, Ana Salgado, David Lindemann, Michael Rosner, Raquel Silva & Onorina Botezat 81
DESIGNING THE MACROSTRUCTURE AND MICROSTRUCTURE OF THE ENEOLI MULTILINGUAL GLOSSARY OF NEOLOGY

Simon Krek, Iztok Kosem & Polona Gantar 85
BRIDGING LEXICOGRAPHY AND NLP: The Architecture and Application of the Digital Dictionary Database of Slovene (DDDS)

Amila Kugic 87
ANALYSIS OF SYNTHETIC CLINICAL NEOLOGISMS: A Corpus-Based Analysis of Large Language Model-Generated Content in the Context of Artificial Intelligence-Assisted Clinical Documentation

Nina Ledinek, Mija Michelizza, Špela Petric Žižić, Domen Krvina, Janoš Ježovnik & Andrej Perdih 89
USING LLMS IN DICTIONARY COMPILATION: The Case of Classifying Adjectives in the Dictionary of the Slovenian Standard Language eSSKJ

Seulki Lee 92
REFRAMING WAN (完), HAO (好), AND DAO (到) IN CHINESE–KOREAN BILINGUAL LEXICOGRAPHY: A Corpus-Based Cognitive Analysis

Magdalena Majdak, Ewa Rodek, Katarzyna Kryńska & Jagoda Marszałek 95
THE ROLE OF LLMS IN THE AUTOMATIC DISTRIBUTION OF MEANINGS FROM KNAPIUSZ'S POLISH–LATIN–GREEK DICTIONARY OF 1643: A Pilot Study

Leonor Martins & Ana Salgado 98
NEOLOGISMS IN EVARISTO.AI VS. ChatGPT: Implications for LLM-Based Lexicography in European Portuguese

Abdelouahab Masbah, Andrea Bellandi & Silvia Piccini 101
DEC FACILE: Semantic-Web Authoring for Combinatorial Explanatory Dictionaries

Nikos Mathioudakis & Alexandra Fiotaki 105
DOCUMENTING THE UNDOCUMENTED: Leveraging LLMs for Neologism Integration into Lexicons

Ondřej Matuška, Miloš Jakubíček & Vojtěch Kovář 108
MANAGEMENT OF LARGE LEXICOGRAPHIC PROJECTS WITH LEXONOMY

Cordula Meißner, Anna-Lena Randermann & Janina Deilke 109
INDEXICALITY IN THE VOCABULARY OF GERMAN COMMUNICATION VERBS: Examining language use and speaker knowledge as a basis for lexicographical conception

Olga Menagarishvili 112
DICTIONARIES OF ENGINEERING IN THE AGE OF AI

Tomasz Michta & Tomasz Czyż 114
CHATGPT OR A DICTIONARY? Resolving polysemy in L2 reception

Tomasz Michta & Katarzyna Mroczyńska 116
SPOILT FOR CHOICE? ChatGPT versus a bilingual dictionary in selecting L2 equivalents for polysemous L1 words

Tomasz Michta, Geraint Paul Rees & Ana Frankenberg-Garcia 118
CORPUS-BASED VERSUS AI-GENERATED DICTIONARY EXAMPLES: How They Differ and What Learners Prefer

Julie Moore 120
DIRECTIONS IN LEXICOGRAPHY IN THE GLOBAL SOUTH

Phoebe Nicholson, Will Rogers & Kate Wild 123
HISTORICAL LEXICOGRAPHY IN THE AGE OF AI: A Progress Report from the OED

Maciej Ogrodniczuk & Dariusz Czerski 125
CAN LLMS AUTOMATE THE NEOLOGISM LIFE CYCLE? Evaluating GenAI in Lexicographical Description of New Polish Words

Iryna Ostapova, Mykyta Yablochkov, Alona Dorozhynska & Iuliia Verbynenko 127
DEVELOPMENT OF A WEB APPLICATION FOR A MULTILINGUAL TERMINOLOGICAL DICTIONARY

Ana Ostroški Anić, Federica Vezzani, Kris Heylen & Ilan Kernerman 130
FROM DETECTION TOSTANDARDIZATION: Practices and Criteria for Terminological Neology

Ana Ostroški Anić, Martina Pavić & Lidija Cvikić 133
TYPES OF DEFINITIONS IN CROATIAN PRIMARY SCHOOL SCIENCE TEXTBOOKS

Andrej Perdih, Janoš Ježovnik & Dejan Gabrovšek 135
IMPROVING DICTIONARY HEADWORD LISTS WITH WORD-PREVALENCE METRICS: An Examination of Low-Prevalence, High-Frequency Words in Slovenian

Silvia Piccini, Andrea Bellandi, Giuliana Elizabeth Vilela Ruiz & Davide Saponaro 138
A SEMANTIC WEB MODEL FOR DIACHRONIC EXPLANATORY COMBINATORIAL DICTIONARIES: Capturing Conceptual Change and Denominational Variation

Geraint Paul Rees, Blanca Arias-Badia, Elisenda Bernal, Marta Brescia-Zapata, Clara Anush Del Rey Adanalian, Ana Frankenberg-Garcia, Marta Garcia Casado, Paula Igareda González, Daniel Jiménez Casas, Robert Lew, Alba Milà Garcia, Guiomar Salvat, Aurora Troncoso Ruiz, Sergi Torner & Patrick Zabalbeascoa 142
ACCESSIBLE DIGITAL LANGUAGE RESOURCES (ADLAR)

Table of Contents

EXTENDED ABSTRACTS

Randy Christian Saputra & Sílvia Barbosa 146
MAPPING EMERGING AI TERMINOLOGY THROUGH AUTOMATICALLY BUILT KNOWLEDGE GRAPHS

Barbara Schmidt-Thieme, Theresa Kruse & Ulrich Heid 149
LEXICAL AND CONCEPTUAL TOOLS FOR LEARNING MATHEMATICS

Nicolai Hartvig Sørensen & Andrea Stengaard 151
BEYOND WORD LISTS – BUT NOT BEYOND DICTIONARIES

Stefania Spina, Irene Fioravanti & Fabio Zanda 155
THE LEARNER DICTIONARY OF ITALIAN COLLOCATIONS (DICI-A): A New Lexicographic Resource

Petra Storjohann 158
DISCURSIVE DYNAMICS OF PEJORATION: The Term *Pushback* between Usage and Lexicography

Maria Tuulik, Kristina Koppel, Ene Vainik, Margit Langemets, Lydia Risberg, Hanna Maask, Esta Prangel & Liina Lutsepp 160
FROM USER FEEDBACK TO BETTER DEFINITIONS: Applying LLMs for Pedagogical Lexicography

Efi Tzelepi & Zoe Gavriilidou 162
DIGITAL LEXICOGRAPHY FOR STEM EDUCATION: Designing a Multilingual Digital Glossary of Scientific Terminology in Greek Sign Language

Perina Vukša Nahod 165
FROM FIELD RECORDINGS TO AN ONLINE DICTIONARY: A Model for Processing Štokavian Dialect Material

Jonas Winnerlöv 167
LEARNER'S DICTIONARIES BASED ON ETYMOLOGICAL PRINCIPLES: The example of a German textbook dictionary for Swedish learners

Sylwia Wojciechowska 169
EXPLORING LARGE LANGUAGE MODELS IN WORD SENSE DISAMBIGUATION: A Study of Corpus Examples for Perception Verbs in learners' dictionaries

Akerke Yessenali 171
PHRASEOLOGICAL THEORETICAL GAPS IN DIGITAL LEXICOGRAPHY: A Contrastive Study of English–German Phraseologisms with Gender Representations

Manjola Zaçellari 174
THE MAKING OF MEANING: From the 1980 Communist Albanian Dictionary to the Great Dictionary of 2026

Tanara Zingano Kuhn, Vojtěch Kovář & Miloš Jakubíček 176
MAPPING SOCIOLINGUISTIC THEORY INTO PRACTICE: Headword candidate selection in the dictionary of pluricentric Portuguese

Valeria Zotti & Rita Gramellini 179
FROM PLURILINGUAL LEXICOGRAPHY AND TERMINOGRAPHY TO GENERATIVE AI TRANSLATION: Potential, Limits, and the Enduring Role of Dictionaries

PART IV SHORT ABSTRACTS

Andrea Abel, Luca Ducceschi & Federico Villa 183
DETECTING SEMANTIC VARIATION BETWEEN GERMAN STANDARD VARIETIES THROUGH WORD EMBEDDING COMPARISON

Luke Omoyemi Akinremi & Elena Martín-Cancela 184
BILINGUAL LEXICOGRAPHY UNDER RESOURCE ASYMMETRY: Designing Yoruba–German dictionary articles in dialogue with Galician–German

Iván Arias-Arias, María José Domínguez Vázquez, María Teresa Sanmarco Bande & Carlos Valcárcel Riveiro 185
FROM CONTENT TO CODE AND BACK: Evaluating Agentic Coding in LLMs for Building R Shiny Dictionary Interfaces

Petra Bago & Ivana Eterović 186
COMPILING A 16TH-CENTURY CROATIAN DICTIONARY IN THE AGE OF AI: Challenges and Perspectives

Martina Bajčić & Ana Ostroški Anić 187
DEFINING LEGAL CONCEPTS IN THE AI AGE: An Exploratory Study into Influencer Practices

Elena Berthemet 188
DESIGNING AN IDIOMATIC DICTIONARY

Harald Bichlmeier 189
VOL. IX OF THE *ETYMOLOGICAL DICTIONARY OF OLD HIGH GERMAN* [*ETYMOLOGISCHES WÖRTERBUCH DES ALTHOCHDEUTSCHEN* (*EWA*)]: History Structure, Contents

Riccardo Billero 190
LEXICOGRAPHY IN THE AGE OF AI: Training Systems with Human Expertise in Art Terminology

Ben Black Bear, Jr., Richard Two Dogs & Jan Ullrich 191
BUILDING THE LARGEST DICTIONARY FOR AN INDIGENOUS LANGUAGE OF NORTH AMERICA

Marek Blahuš, Miloš Jakubíček, Katarína Petreková & Yingtian Wu 192
AUTOMATIC DRAFTING OF TERMINOLOGICAL DICTIONARIES WITH ONE-CLICK DICTIONARY FOR 34 LANGUAGES

Igor Boguslavsky, Vyacheslav Dikonov, Evgeniya Inshakova, Alexandre Lazursky, Svetlana Timoshenko & Tatiana Frolova 193
LEXICOGRAPHIC RESOURCES FOR INFERENCE-BASED SEMANTIC TEXT PROCESSING SYSTEM

Ivana Brač 195
SEMANTIC CLASS ASSIGNMENT IN A CROATIAN VERB LEXICON: Challenges of Applying VerbNet's Classes

Katja Brankačkec & Barbora Štěpánková 196
MULTIFUNCTIONAL FUNCTION WORDS IN THE UPPER SORBIAN-CZECH DICTIONARY MUDRA 2.0

Cormac Breathnach & Pádraig Ó Mianáin 197
REVERSING THE NEW ENGLISH–IRISH DICTIONARY DATABASE TO CREATE ENTRY FRAMEWORKS FOR IRISH–ENGLISH AND MONOLINGUAL IRISH DICTIONARIES

Table of Contents

SHORT ABSTRACTS

Rukiye Smilla Burkart & Susanne Kabatnik 198
WIE DEFINIERT MAN DAS UNBEKANNTE? Prompt Engineering als Methode der KI-gestützten Bedeutungsdefinition im Pandemictionary

Valeria Caruso & Geraint Paul Rees 199
FROM LEXICAL THEORY TO AI TRANSLATION: Procomplement Verbs and ChatGPT

Lian Chen, Damien Nouvel, Huy-Linh Dao & Alexander Delaporte 200
FROM CORPUS TO DICTIONARY: OntoLex-Lemon-based modelling of neology in Chinese–French and Vietnamese–French computational lexicography

Luisa Cimander 201
ENGLISH ADJECTIVES IN GERMAN: Processes of Integration since the 17th Century and Lexicographic Implications

Li Cui, Seol Namkung, Jun Lee, Hae-Yun Jung & Kilim Nam 202
WHAT LLMS LEARN AND WHAT THEY MISS: Frequency Effects in the Automated Description of Korean Verbal Inflection

Flávia de Oliveira Maia-Pires & Silvio Licurgo Pires 203
USER-ORIENTED SPECIALIZED DIGITAL LEXICOGRAPHY AND TERMINOLOGICAL ACCESSIBILITY: The LexiC platform applied to the glossary of Brazilian migration legislation

Gilles-Maurice de Schryver 204
CHATBOT LEXICOGRAPHY: Are GenAI LLMs simply 'dictionaries' in their own right and 'lexicographers' plain and simple?

Gilles-Maurice de Schryver 205
THE NINTH ART IN LEXICOGRAPHY: The power of comic strips

Ida Dringó-Horváth & Katalin P. Márkus 206
DICTIONARY USE IN A TECHNOLOGY-DRIVEN ERA: Longitudinal Evidence from EFL and GFL Graduates

Sophia Fiedler 207
TOWARDS AN INTERACTIONAL LEXICON OF SPOKEN GERMAN: A Functional, Audio-Based Lexicographic Resource

Ivana Filipović Petrović & Slobodan Beliga 208
MODELLING PHRASEOLOGICAL SEMANTIC STRUCTURES IN DMLEX: Aligning a Croatian Lexonomy-Based Idiom Dictionary with a Thematic Index

Hanna Fischer, Alfred Lameli, Nathalie Mederake & Nico Urbach 209
LINKING DICTIONARIES THROUGH MEANING: A sense-level evaluation of taxonomy-guided LLM classification in lexicography

Margarita Freixas, Esther Hernández & Marija Žarković Eriksson 210
E-*DRAE* 1884: Towards a Digital and Hypertextual Critical of a Nineteenth-Century Dictionary

Christine Ganslmayer & Martina Nied Curcio 211
HOW GERMAN LANGUAGE TEXTBOOKS ADDRESS THE DIGITAL TRANSFORMATION OF LEXICOGRAPHY

Marta Garcia Casado 212
NEW SENSES, NEW METHODS: Clustering as a guide for detecting semantic neologisms

Pol Garriga Martínez 213
THE VOCABULARY OF MENTAL HEALTH: Exploitation of specialized diachronic corpora with artificial intelligence for qualitative semantic analysis

Zoe Gavriilidou 214
SPANISH LOANWORDS IN MODERN GREEK: Linguistic Analysis and Lexicographic Treatment in the Dictionary of Standard Modern Greek

Zoe Gavriilidou & Evanthia Konstantinidou 215
PEDAGOGICAL LEXICOGRAPHY IN PRACTICE: Raising Dictionary Use Strategies through Card-Based Instruction

Alexander Geyken & Gregor Middell 216
TREND WORD DETECTION FOR DWDS, A LEXICAL INFORMATION SYSTEM FOR CONTEMPORARY GERMAN

Rufus H. Gouws & Theo J.D. Bothma 217
OPTIMISING INTERFACE DESIGN IN ONLINE LEXICOGRAPHIC PRODUCTS

Barbara Heinisch 218
DICTIONARIES, GLOSSARIES AND TERMINOLOGY DATABASES IN THE AGE OF ARTIFICIAL INTELLIGENCE: Rethinking boundaries and functions

Ondřej Herman 219
SENSE CLASSIFICATION FOR WORD SKETCHES

Polina Iaroshenko & Natalia Loukachevitch 220
MORPHOLOGICAL FORMS IN EMOTION LEXICONS: Evidence from LLMs

Ellert Thor Johannsson, Jóhannes B. Sigtryggsson & Thordis Ulfarsdottir 221
UNLOCKING A NINETEENTH CENTURY DICTIONARY: A Closer Look at Konráð Gíslason's *Dönsk orðabók* (1851)

Boris Kern 222
WORD-FORMATION CHAINS IN SLOVENIAN: Evaluating the Effectiveness of AI

Václava Kettnerová, Veronika Kolářová & Jiří Mírovský 223
WHERE THE BOUNDARY BETWEEN DERIVATIONAL AND INFLECTIONAL MORPHOLOGY MEETS LEXICOGRAPHY: Building the DeriVallex Lexicon

Agnes Kim 224
DIGITAL METALEXICOGRAPHY AS A TOOL FOR HISTORICAL (CONTACT) LINGUISTICS

Veronika Kolářová, Václava Kettnerová, Michal Olbrich & Jiří Mírovský 225
NOMVALLEX: Valency of Czech nouns and adjectives in networks of derivationally related words

Mariia Kopylova 226
AI-ASSISTED DETECTION AND MITIGATION OF LEXICOGRAPHIC BIAS: Semantic Reduction in Definitions of Five Socio-Legal Terms

Katharina Korecky-Kröll, Philipp Stöckle, Daniel Elsner, Wolfgang Koppensteiner & Elisabeth Eder 227
LLM-ASSISTED LEXICOGRAPHY IN LOW-RESOURCE CONTEXTS: From prompt to dictionary article

Iztok Kosem, Polona Gantar, Simon Krek & Tjaša Arčon 228
DIVIDE AND DEFINE WITH LLMS: Generating Definitions for the Digital Dictionary Database of Slovene

SHORT ABSTRACTS

František Kovařík, Marek Blahuš, Miloš Jakubíček & Vojtěch Kovář 229
AGREE TO DISAGREE: Intra-annotator Agreement in Semi-Automatic Vocabulary Building

Dominika Kováříková & Laura A. Janda 230
EMOTIONS IN CONTEXT: A Semantic Field Viewed through Grammar and Collocations

Barbara Lewandowska-Tomaszczyk 231
THE INTEGRATION OF BRITISH AND AMERICAN SLANG IN CONTEMPORARY POLISH AND THE ROLE OF LLM-DRIVEN DYNAMIC LEXICOGRAPHY

David Lindemann 232
DEVELOPING A DIGITAL VERSION OF GIDOR BILBAO'S LATIN-BASQUE DICTIONARY

Ivana Lovrić Jović & Lana Hudeček 233
SYNONYMS THAT NEVER MET: On the Compilation of the Dictionary of the Dubrovnik Idiom

Tinatin Margalitadze, Giorgi Meladze, Ketevan Mchedlishvili, Tamar Laluashvili & Giorgi Okropiridze 234
EFFECTIVENESS OF AI FOR THE LEXICOGRAPHY OF LOW-RESOURCE LANGUAGES: The Dictionary of Georgian Neologisms

Aleksandra Marković, Cvetana Krstev, Milica Dinić Marinković & Ranka Stanković 236
TOWARDS A REPOSITORY OF SIMILES IN CONTEMPORARY SERBIAN

Peter Meyer 237
AN AI-DRIVEN NATURAL LANGUAGE INTERFACE FOR GRAPH DATABASE QUERIES IN A LEXICOGRAPHICAL RESOURCE

Ana Mihaljević & Josip Mihaljević 238
ARTIFICIAL INTELLIGENCE AS SUPPORT FOR THE DIGITIZATION AND COMPILATION OF THE *DICTIONARY OF THE CROATIAN REDACTION OF CHURCH SLAVONIC*

Tibor M. Pintér, Manana Rusieshvili-Cartledge, Dóra Pődör, Katalin P. Márkus & Marine Makhatadze 239
SUBLEXICAL SYSTEMS OF COLOUR IN BIBLICAL GEORGIAN AND HUNGARIAN: Findings of the Pentateuch

Bruno Nahod & Marijana Tuta Dujmović 240
FROM TEXT TO TERMS: TermScout, an AI Module for Early Terminology Processing

Gunta Nešpore-Bērzkalne, Mikus Grasmanis & Madara Stāde 241
POSSIBILITIES OF REPRESENTATION OF MULTI-WORD EXPRESSIONS IN THE LATVIAN ELECTRONIC DICTIONARY *TĒZAURS*

Sanni Nimb, Ida Flörke, Agnes Aggergaard Mikkelsen, Nathalie Hau Norman & Nicolai Hartvig Sørensen 242
AI USED TO DETECT PROBLEMATIC CONTENT IN CORPUS EXAMPLES

Krzysztof Nowak, Iwona Krawczyk & Jagoda Marszałek 243
DIALEXPOL: LLM-assisted Sense Assignment for a Diachronic Dictionary of Latin

Krzysztof Nowak & Dorota Mika 244
POLILEX: MODELLING HETEROGENEOUS LEXICAL RESOURCES:
Lessons from the Dictionary of Polish Dialects

Emma Josefin Ölander Aadland 245
A CRITICAL LEXICOGRAPHICAL DISCOURSE ANALYSIS OF NEW ADJECTIVAL ENTRIES IN THE STANDARD DICTIONARIES FOR NORWEGIAN

Benedikt Perak, Dragana Špica & Irena Srdanović 246
A GRAPH-BASED LEXICOGRAPHIC WORKFLOW FOR JAPANESE ADJECTIVES AND THEIR LEXICAL RELATIONS IN AN AI TUTOR

Linda Prossliner 247
AI-SUPPORTED PHRASEOGRAPHY FOR YOUNG LEARNERS: Designing Definitions and Illustrations through AI – Opportunities and Challenges

Laura Rebosio 248
OSTENSIVE AIDS IN LEARNER'S DICTIONARIES: A diachronic, comparative analysis

Lydia Risberg, Kristina Koppel, Margit Langemets, Hanna Maask, Esta Prangel, Maria Tuulik & Silver Vapper 249
ARE BEARS MORE OFFENSIVE THAN DOGS WHEN DRUNK? Using LLMs to systematize the labelling of Estonian synonyms

Ines Röhrer & Manuel Raaf 251
REFLECTIONS ON THE SEMANTIC UNDERSTANDING OF GENERATIVE AI:
From LLM experiments to lexicographical tools

Antonio San Martín & Catherine Trekker 252
INTRODUCING HUMAN-CENTEREDNESS IN AI-ASSISTED LEXICOGRAPHY

Saghar Sharifi & Abolfazl Alamdar 253
TOWARDS A HANDBOOK OF PERSIAN LEXICOGRAPHY:
A Lexicographical-Functions Analysis of 20th–21st Century Dictionaries

Jóhannes B. Sigtryggsson 254
THE EFFECTS OF PURISM ON A NINETEENTH-CENTURY ICELANDIC BILINGUAL DICTIONARY

Hana Skoumalová 255
EXPANDING THE MWE DATABASE WITH VARIANTS

Deepak Solanki 256
USER-ORIENTED SPECIALIZED LEXICOGRAPHY: Corpus-Driven Models for Targeted Dictionary Design

Michael Spagnol, John J. Camilleri & Dwayne Ellul 257
FROM *ḊAMMA* TO *ĠABRA*: Four Centuries of Maltese Lexicography

Vasyl Starko & Andriy Rysin 258
LARGE-SCALE SEMIAUTOMATIC DETECTION OF NEW AND UNREGISTERED WORDS IN MODERN UKRAINIAN

Nika Stefan, Johan van der Zwaag & Hindrik Sijens 259
BEYOND THE SINGLE STANDARD: Addressing Variation and User Needs in an Online Frisian Dictionary

Petra Storjohann 260
THE INTERPLAY BETWEEN INDIVIDUAL NEW LEXICAL ITEMS AND MULTI-LAYERED DISCOURSE REPRESENTATIONS OF NEOLOGISMS

SHORT ABSTRACTS

Silga Sviķe 261
A LEXICOGRAPHICAL ANALYSIS OF NEOLOGISM DICTIONARIES BASED ON THE MULTILINGUAL DICTIONARY OF NEW WORDS (2010–2024)

Pius ten Hacken 262
THE THEORETICAL FOUNDATION OF DICTIONARY DEFINITIONS

Dušica Terzić, Dejana Mirković-Birtašić & Saša Marjanović 263
TOWARDS A FRENCH–SERBIAN DICTIONARY OF SPORTS TERMS: Corpus-based or AI-Assisted Methodology?

Carole Tiberius, Martin Kroon, Lut Colman & Dalí Chirino 264
RANKING COLLOCATIONS IN *WOORDCOMBINATIES*

Jesús Torres del Rey & M.ª Teresa Fuentes Morán 265
A HYBRID APPROACH TO WEB ACCESSIBILITY EVALUATION IN DIGITAL LEXICOGRAPHY: Comparing and Integrating Rule-Based Validators and Generative AI

Ágoston Tóth 266
(UG)TPEX: A (Usage-Guided) Typical Example Score

Ágoston Tóth & Esra Abdelzaher 267
GPT AS A SENSE-TAGGER FOR SELECTING DICTIONARY EXAMPLES

Lars Trap-Jensen & Henrik Lorentzen 268
REVISING THE USAGE-LABELLING SYSTEM OF THE DANISH DICTIONARY

Thordis Ulfarsdottir & Ellert Thor Johannsson 269
THE REVERSAL OF A BILINGUAL DICTIONARY USING AI

Anna Vacalopoulou 270
A DEFINITION WALKS INTO A DICTIONARY: The lexicography of humor

Ene Vainik, Heete Sahkai, Ahto Kiil, Jelena Kallas & Geda Paulsen 271
SEMANTIC CLUSTERING AND CATEGORISATION OF CONSTRUCTIONAL COLLEXEMES USING LLMS

Carlos Valcárcel Riveiro, María Teresa Sanmarco Bande, Iván Arias-Arias & María José Domínguez Vázquez 272
CAN LEXICOGRAPHIC ENTRIES EMBED DIDACTIC MODULES? A Case Study on LLM-Generated Outputs for a Dictionary Interface

Martina Vokáčová & Anna Marklová 273
AFFORDANCE NORMS IN CZECH AND ENGLISH: Human–LLM Misalignment in Embodied Lexical Knowledge

Zongyuan Wang, Valeria Caruso & Anna De Meo 274
COLLABORATIVE LEXICOGRAPHY FOR HERITAGE LANGUAGE LEARNING

Tarrin Wills & Ellert Thor Johannsson 275
CAN AI-ASSISTED NORMALISATION IMPROVE ACCESS TO HISTORICAL LEXICOGRAPHIC DATA?

Shigeru Yamada 276
PEDAGOGICAL VALUE AND INSIGHTS FROM COMPARING EFL DICTIONARIES: Evidence from *apparently*, *be on speaking terms*, and *suffer* (*from*)

Yelena Yerznkyan & Diana Movsisyan 277
ENCODING EMOTIVITY IN DICTIONARY DEFINITIONS: A Corpus-Based Study of Verbs of Understanding

Marija Žarković Eriksson 278
LEGAL LEXICON AND THE E-*DRAE* 1884: A Diachronic Study of Specialised Vocabulary in a General Dictionary Using a Digital Critical Edition

PART V WORKSHOPS

Ilan Kernerman & Kris Heylen 280
8TH GLOBALEX WORKSHOP ON LEXICOGRAPHY AND NEOLOGY

Voula Giouli, Athina Sioupi, Nina Böbel & Alexander Ziem 282
CONSTRUCTICOGRAPHY IN PRACTICE: Constructions, Challenges, and Opportunities

Ondřej Matuška 284
POSTEDITING LEXICOGRAPHY WITH SKETCH ENGINE AND LEXONOMY

Lian Chen & Besim Kabashi 285
MULTI-WORD EXPRESSIONS AND PHRASEOLOGY: Corpus-based and computer-processed

INDEX OF AUTHORS

List of Authors 290

Index of Authors 304

ACKNOWLEDGEMENTS

We want to express our sincere gratitude to all sponsors who generously supported the XXII EURALEX International Congress financially. Your contributions were invaluable, and we greatly appreciate your commitment to the success of the event.

Sponsors

CLARIAH-AT

Stadt Wien

Hornby
A. S. Hornby Educational Trust CIO

SKETCH ENGINE

We would like to sincerely thank all those who invested their time and expertise in reviewing the submissions. Your invaluable work has significantly enhanced the quality of our conference, and we greatly appreciate your commitment and support.

Programme Committee

- Carolina Flinz, University of Pavia
- Wolfgang Koppensteiner, Austrian Academy of Sciences
- Katharina Korecky-Kröll, Austrian Academy of Sciences
- Iztok Kosem, Faculty of Arts, University of Ljubljana
- Veronika Lipp, ELTE Research Centre for Linguistics
- Geraint Paul Rees, Universitat Pompeu Fabra
- Philipp Stöckle, Austrian Academy of Sciences
- Kristina Strkalj Despot, Institute of Croatian Language and Linguistics
- Tanja Wissik, Austrian Academy of Sciences

Scientific Committee

- Andrea Abel, European Academy Bolzano/Bozen
- Arleta Adamska-Sałaciak, Adam Mickiewicz University, Poznan
- Špela Arhar Holdt, Centre for Language Resources and Technologies (CJVT), University of Ljubljana
- Petra Bago, Faculty of Humanities and Social Sciences, University of Zagreb
- Martina Bajcic, Faculty of Law, University of Rijeka
- Vincent Balnat, University of Strasbourg
- Verginica Barbu Mititelu, Romanian Academy Research Institute for Artificial Intelligence
- Slobodan Beliga, Department of Informatics, University of Rijeka
- Vladimír Benko, Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences
- Úna Bhreathnach, Fiontar & Scoil na Gaeilge, Dublin City University
- Hans Bickel, Schweizerisches Idiotikon
- Goranka Blagus Bartolec, Institute of Croatian Language and Linguistics
- Petar Božović, University of Montenegro
- Ivana Brač, Institute of Croatian Language and Linguistics
- Maja Bratanić, Institute of Croatian Language and Linguistics
- Thomas Burch, Trier Center for Digital Humanities
- Valeria Caruso, Università degli Studi di Napoli „l'Orientale"
- Lian Chen, LLL, University of Orléans
- Rute Costa, Universidade NOVA de Lisboa
- Janet Decesaris, Universitat Pompeu Fabra
- Idalete Dias, University of Minho
- María José Domínguez Vázquez, Universidade de Santiago de Compostela

SCIENTIFIC COMMITTEE

- Amelie Dorn, Austrian Academy of Sciences
- Anna Dziemianko, Adam Mickiewicz University
- Pamela Faber, University of Granada
- Ilse Feinauer, University of Stellenbosch
- Ivana Filipović Petrović, Croatian Academy of Sciences
- Edward Finegan, University of Southern California
- Fabian Fleißner, Austrian Academy of Sciences
- Carolina Flinz, University of Milan
- Thierry Fontenelle, European Investment Bank
- Judit Freixa, Universitat Pompeu Fabra
- Polona Gantar, University of Ljubljana
- Radovan Garabík, Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences
- Zoe Gavriilidou, Democritus University of Thrace
- Andreas Gellan, Austrian Academy of Sciences
- Alexander Geyken, Berlin-Brandenburg Academy of Sciences and Humanities
- Voula Giouli, Aristotle University of Thessaloniki
- Volker Harm, Göttingen Academy of Sciences and Humanities
- Kris Heylen, Dutch Language Institute (INT, Leiden) / KU Leuven (University of Leuven), Research Unit QLVL
- Lana Hudeček, Institute of Croatian Language and Linguistics
- Max Ionov, University of Zaragoza
- Dubravka Ivšić Majić, Institute of Croatian Language
- Miloš Jakubíček, Lexical Computing
- Besim Kabashi, University of Tübingen
- Jelena Kallas, Institute of the Estonian Language
- Virna Karlić, Faculty of Humanities and Social Sciences, University of Zagreb
- Boris Kern, Fran Ramovš Institute of the Slovenian Language
- Ilan Kernerman, K Dictionaries
- Fahad Khan, Istituto di Linguistica Computazionale „Antonio Zampolli“
- Svetla Koeva, Institute for Bulgarian Language
- Veronika Kolářová, Institute of Formal and Applied Linguistics, Charles University, Faculty of Mathematics and Physics
- Kristina Koppel, Institute of the Estonian Language
- Wolfgang Koppensteiner, Austrian Academy of Sciences
- Katharina Korecky-Kröll, Austrian Academy of Sciences
- Iztok Kosem, Faculty of Arts, University of Ljubljana
- Maria Luisa Krapp, Austrian Academy of Sciences
- Simon Krek, Jožef Stefan Institute
- Margit Langemets, Institute of the Estonian Language

SCIENTIFIC COMMITTEE

- Alexandra N. Lenz, Austrian Academy of Sciences
- Marie-Claude L'Homme, Observatoire de linguistique Sens-Texte
- David Lindemann, University of the Basque Country
- Veronika Lipp, ELTE Research Centre for Linguistics
- Anja Lobenstein-Reichmann, Göttingen Academy of Sciences and Humanities
- Irina Lobzhanidze, Ilia State University
- Henrik Lorentzen, Society for Danish Language and Literature
- Magdalena Majdak, Institute of Polish Language, Polish Academy of Sciences
- Tinatin Margalitadze, Ilia State University
- Saša Marjanović, Faculty of Philology, University of Belgrade
- Ivana Matas Ivanković, Institute of Croatian Language and Linguistics
- Peter Meyer, Leibniz Institute for the German Language (IDS) Mannheim
- Milica Mihaljević, Institute for Croatian Language and Linguistics
- Karlheinz Moerth, Austrian Academy of Sciences
- Fabio Mollica, University of Milan
- Hilary Nesi, Coventry University
- Henrik Nilsson, European Association for Terminology, C.A.G
- Pär Nilsson, Swedish Academy Dictionary
- Vincent Ooi, National University of Singapore
- Christian-Emil Ore, University of Oslo
- Petya Osenova, Sofia University / Institute of Information and Communication Technologies Bulgarian Academy of Sciences
- Ana Ostroški Anić, Institute for the Croatian Language and Linguistics
- Gokhan Ozkan, Bogazici University
- Martina Pavić, Institute for Croatian Language and Linguistics
- Sanja Perić Gavrančić, Institute for Croatian Language and Linguistics
- Ralf Plate, Academy of Sciences and Literature Mainz
- Boris Pritchard, University of Rijeka
- Natascia Ralli, Institute for Applied Research, Eurac Research
- Geraint Paul Rees, Universitat Rovira i Virgili
- Irene Renau, Pontificia Universidad Católica de Valparaíso
- Ana Salgado, Universidade NOVA de Lisboa
- Hindrik Sijens, Fryske Akademy
- Ranka Stanković, University of Belgrade
- Egon W. Stemle, Eurac Research
- Frieda Steurs, Faculty of Arts, KU Leuven
- Philipp Stöckle, Austrian Academy of Sciences
- Kristina Strkalj Despot, Institute of Croatian Language and Linguistics

SCIENTIFIC COMMITTEE

- Elsabe Taljard, University of Pretoria
- Giovanni Tallarico, University of Verona
- Pius ten Hacken, University of Innsbruck
- Carole Tiberius, Instituut voor de Nederlandse Taal / Leiden University Centre for Linguistics
- Lars Trap-Jensen, Society for Danish Language and Literature
- Anna Vacalopoulou, Athena RC (ILSP)
- Carlos Valcárcel Riveiro, Universidade de Vigo
- Federica Vezzani, University of Padua
- Craig Volker, James Cook University
- Tanja Wissik, Austrian Academy of Sciences
- Sascha Wolfer, Leibniz Institute for the German Language (IDS) Mannheim
- Mojca Žagar Karer, Research Centre of the Slovenian Academy of Sciences and Arts (ZRC SAZU)
- Alexander Ziem, Heinrich Heine University Düsseldorf
- Tanara Zingano Kuhn, University of Coimbra

Special thanks

We also extend our special thanks to Leona Fischer, Hannah Grischany, Magdalena Kirnbauer, Simone Sailer, Sonja Schwaiger and Anja Wittibschlager for their careful support in ensuring that this book meets all formal requirements.

PART I

OVERVIEW

EURALEX 2026

OVERVIEW

Keynotes:

- *Robert Lew*
  With Bots Like These, Who Needs Dictionaries?
- *Kate Wild*
  Beyond the Entry: The Evolution of Historical Lexicography at the Oxford English Dictionary
- *Andreas Baumann*
  Polysemy: Operationalization, Emergence, Perception, and the Human in the Loop
- *Ana Salgado*
  Dictionaries as Pipelines: Lexicographic Work at the Lisbon Academy of Sciences
- *Jan Hajič*
  Large Language Models, Symbolic Meaning Representations and Ontologies: Perspectives and Questions

Extended Abstracts:

1. *Vassiliki Afentoulidou, Ioannis Tziogkas, Christos Papanagiotou, Aretousa Giannakou & Anastasia Christofidou*
   AI-Assisted Lexicography in Modern Greek: Neology and Genre
2. *Mikel Alonso*
   AI-Assisted Information Extraction from Larramendi's Trilingual Dictionary (1745)
3. *Maria Ammari*
   Pedagogy as the Core of Education: A Domain-Analytic Route to a Corpus-based Dictionary
4. *Michela Bandini, Silvia Piccini, Andrea Bellandi & Emiliano Giovannetti*
   Toward a Narrative-Oriented Dictionary: Computational Modelling of Youth Language in Socially Vulnerable Contexts
5. *Biljana Božinovski, Makimi Kano, Peter Gobel & Mihaela Franjić*
   Lexicographic Functions without Dictionaries: GenAI-Mediated Lexical Problem-Solving Among Non-English-Major Tertiary EFL Students
6. *Gabriela Brůhová & Veronika Raušová*
   The "Hell Of A" Construction in Casual Written English Online: A Synchronic Functional Account
7. *Edmond Cane*
   Learning Like Humans, Parsing Unlike Machines: A Construction-Based Approach to Lexical Annotation
8. *Elena Castaldi*
   Inflected-Form Search in Ancient Greek Dictionaries: A Form-To-Lemma Access Model
9. *Elina Chadjipapa, Zoe Gavriilidou, Maria Mitsiaki & Ifigeneia Dosi*
   Chatbots in Pedagogical Lexicography: Evaluating AI-generated dictionary entries from three LLMs
10. *John J. Chew III, Emma Ferrett & Anastasia Riehl*
    Canadian English Dictionary: A Progress Report
11. *Luisa Cimander & Petra Storjohann*
    Tracing Neologisms in Neo-Discourses: Diachronic and Cross-Domain Perspectives
12. *Václav Cvrček, Martin Stluka & Klára Pivoňková*
    Tracking Eight Centuries: The Czech Monitor Corpus as a Resource for Diachronic Lexical Research

13. *Veronique De Tier, Katrien Depuydt, Koen Mertens & Mathieu Fannee*
From Pilot to Practice: Two Years of Progress in Developing a Shared Infrastructure for Dutch Dialect Dictionaries

14. *Kristina Š. Despot, Ana Ostroški Anić & Matej Klemen*
Beyond Binary Metaphor Detection: Dataset-Driven Evaluation of Conceptual Metaphor Type Recognition in Croatian and Slovene

15. *Emma Ferrett, Amir Ghorvei & Anastasia Riehl*
A Dictionary for All Canadians: Establishing an Inclusive Standard for the Canadian English Dictionary

16. *Emmanuel Sylvain Fomat*
Camfranglais in the Age of Artificial Intelligence: Lexicographic Challenges, Opportunities and a Human-AI Approach to Documenting Linguistic Hybridity in Multilingual Cameroon

17. *Daniele Franceschi*
Gender Representation and Editorial Responsibility in the OED: A case study of *female*, *feminine*, *male* and *masculine*

18. *Anne-Kathrin Gärtig-Bressan & Pia Carmela Lombardi*
Bilingual Dictionaries with Italian in the 19th and 20th Centuries: The ALON portal

19. *Marius Glebus*
Choosing Optimal AI Prompt Instructions for Neologism Detection

20. *Wanjing Han & Gilles-Maurice de Schryver*
Can LLMs Capture China's Emerging Economic Neoconcepts Lexicographically?

21. *Wanjing Han & Gilles-Maurice de Schryver*
Corpus-Based and AI-Assisted Design of a Chinese–English Business Dictionary

22. *Kris Heylen, Katrien Depuydt, Carole Tiberius & Jesse de Does*
Building a Sense Inventory for the Central Lexicographic Infrastructure of Dutch

23. *Kris Heylen, Ilan Kernerman, Judit Freixa, Tarja Heinonen, Hae-Yun Jung, Henrik Lorentzen, Kilim Nam, Benjamin Rouxel, Ana Salgado, Miguel Sánchez Ibáñez, Petra Storjohann & Lars Trap-Jensen*
The ENEOLI Matrix of Neologism Selection Criteria: A Comparative Framework for Lexicographic Practice

24. *Louise Holmer*
Neology Mills? AI-Influenced Newspaper Articles and their Implications for Lexicology and Lexicography

25. *Željko Jozić, Marijana Horvat & Martina Horvat*
Digitization of Adam Patačić's Manuscript Dictionary

26. *Hae-Yun Jung, Jinsan An, Kilim Nam & Jun Lee*
From Neologism Collection to Lexicographic Description: A Corpus-LLM-Lexicographer Triangulation Framework

27. *Ilan Kernerman, Federica Vezzani, Anna Vacalopoulou, John Humbley, Ana Salgado, David Lindemann, Michael Rosner, Raquel Silva & Onorina Botezat*
Designing the Macrostructure and Microstructure of the ENEOLI Multilingual Glossary of Neology

28. *Simon Krek, Iztok Kosem & Polona Gantar*
Bridging Lexicography and NLP: The Architecture and Application of the Digital Dictionary Database of Slovene (DDDS)

29. *Amila Kugic*
Analysis of Synthetic Clinical Neologisms: A Corpus-Based Analysis of Large Language Model-Generated Content in the Context of Artificial Intelligence-Assisted Clinical Documentation

EXTENDED ABSTRACTS

30. *Nina Ledinek, Mija Michelizza, Špela Petric Žižić, Domen Krvina, Janoš Ježovnik & Andrej Perdih*
Using LLMs in Dictionary Compilation: The Case of Classifying Adjectives in the Dictionary of the Slovenian Standard Language eSSKJ

31. *Seulki Lee*
Reframing Wan (完), Hao (好), and Dao (到) in Chinese–Korean Bilingual Lexicography: A Corpus-Based Cognitive Analysis

32. *Magdalena Majdak, Ewa Rodek, Katarzyna Kryńska & Jagoda Marszałek*
The Role of LLMs in the Automatic Distribution of Meanings from Knapiusz's Polish–Latin–Greek Dictionary of 1643: A Pilot Study

33. *Leonor Martins & Ana Salgado*
Neologisms in Evaristo.ai vs. ChatGPT: Implications for LLM-Based Lexicography in European Portuguese

34. *Abdelouahab Masbah, Andrea Bellandi & Silvia Piccini*
DEC Facile: Semantic-Web Authoring for Combinatorial Explanatory Dictionaries

35. *Nikos Mathioudakis & Alexandra Fiotaki*
Documenting the Undocumented: Leveraging LLMs for Neologism Integration into Lexicons

36. *Ondřej Matuška, Miloš Jakubíček & Vojtěch Kovář*
Management of Large Lexicographic Projects with Lexonomy

37. *Cordula Meißner, Anna-Lena Randermann & Janina Deilke*
Indexicality in the Vocabulary of German Communication Verbs: Examining language use and speaker knowledge as a basis for lexicographical conception

38. *Olga Menagarishvili*
DICTIONARIES OF ENGINEERING IN THE AGE OF AI

39. *Tomasz Michta & Tomasz Czyż*
ChatGPT or a Dictionary? Resolving polysemy in L2 reception

40. *Tomasz Michta & Katarzyna Mroczyńska*
Spoilt for Choice? ChatGPT versus a bilingual dictionary in selecting L2 equivalents for polysemous L1 words

41. *Tomasz Michta, Geraint Paul Rees & Ana Frankenberg-Garcia*
Corpus-Based versus AI-Generated Dictionary Examples: How They Differ and What Learners Prefer

42. *Julie Moore*
Directions in Lexicography in the Global South

43. *Phoebe Nicholson, Will Rogers & Kate Wild*
Historical Lexicography in the Age of AI: A Progress Report from the OED

44. *Maciej Ogrodniczuk & Dariusz Czerski*
Can LLMs Automate the Neologism Life Cycle? Evaluating GenAI in Lexicographical Description of New Polish Words

45. *Iryna Ostapova, Mykyta Yablochkov, Alona Dorozhynska & Iuliia Verbynenko*
Development of a Web Application for a Multilingual Terminological Dictionary

46. *Ana Ostroški Anić, Federica Vezzani, Kris Heylen & Ilan Kernerman*
From Detection TOSTANDARDIZATION: Practices and Criteria for Terminological Neology

47. *Ana Ostroški Anić, Martina Pavić & Lidija Cvikić*
Types of Definitions in Croatian Primary School Science Textbooks

48. *Andrej Perdih, Janoš Ježovnik & Dejan Gabrovšek*
Improving Dictionary Headword Lists with Word-Prevalence Metrics: An Examination of Low-Prevalence, High-Frequency Words in Slovenian

49. *Silvia Piccini, Andrea Bellandi, Giuliana Elizabeth Vilela Ruiz & Davide Saponaro*
A Semantic Web Model for Diachronic Explanatory Combinatorial Dictionaries: Capturing Conceptual Change and Denominational Variation

50. *Geraint Paul Rees, Blanca Arias-Badia, Elisenda Bernal, Marta Brescia-Zapata, Clara Anush Del Rey Adanalian, Ana Frankenberg-Garcia, Marta Garcia Casado, Paula Igareda González, Daniel Jiménez Casas, Robert Lew, Alba Milà Garcia, Guiomar Salvat, Aurora Troncoso Ruiz, Sergi Torner & Patrick Zabalbeascoa*
Accessible Digital Language Resources (ADLAR)

51. *Randy Christian Saputra & Sílvia Barbosa*
Mapping Emerging AI Terminology Through Automatically Built Knowledge Graphs

52. *Barbara Schmidt-Thieme, Theresa Kruse & Ulrich Heid*
Lexical and Conceptual Tools for Learning Mathematics

53. *Nicolai Hartvig Sørensen & Andrea Stengaard*
Beyond Word Lists – But Not Beyond Dictionaries

54. *Stefania Spina, Irene Fioravanti & Fabio Zanda*
The Learner Dictionary of Italian Collocations (DICI-A): A New Lexicographic Resource

55. *Petra Storjohann*
Discursive Dynamics of Pejoration: The Term *Pushback* between Usage and Lexicography

56. *Maria Tuulik, Kristina Koppel, Ene Vainik, Margit Langemets, Lydia Risberg, Hanna Maask, Esta Prangel & Liina Lutsepp*
From User Feedback to Better Definitions: Applying LLMs for Pedagogical Lexicography

57. *Efi Tzelepi & Zoe Gavriilidou*
Digital Lexicography for STEM Education: Designing a Multilingual Digital Glossary of Scientific Terminology in Greek Sign Language

58. *Perina Vukša Nahod*
From Field Recordings to an Online Dictionary: A Model for Processing Štokavian Dialect Material

59. *Jonas Winnerlöv*
Learner's Dictionaries Based on Etymological Principles: The example of a German textbook dictionary for Swedish learners

60. *Sylwia Wojciechowska*
Exploring Large Language Models in Word Sense Disambiguation: A Study of Corpus Examples for Perception Verbs in learners' dictionaries

61. *Akerke Yessenali*
Phraseological Theoretical Gaps in Digital Lexicography: A Contrastive Study of English–German Phraseologisms with Gender Representations

62. *Manjola Zaçellari*
The Making of Meaning: From the 1980 Communist Albanian Dictionary to the Great Dictionary of 2026

63. *Tanara Zingano Kuhn, Vojtěch Kovář & Miloš Jakubíček*
Mapping Sociolinguistic Theory into Practice: Headword candidate selection in the dictionary of pluricentric Portuguese

64. *Valeria Zotti & Rita Gramellini*
From Plurilingual Lexicography and Terminography to Generative AI Translation: Potential, Limits, and the Enduring Role of Dictionaries

Short Abstracts:

1. *Andrea Abel, Luca Ducceschi & Federico Villa*
   Detecting Semantic Variation between German Standard Varieties through Word Embedding Comparison
2. *Luke Omoyemi Akinremi & Elena Martín-Cancela*
   Bilingual Lexicography Under Resource Asymmetry: Designing Yoruba–German dictionary articles in dialogue with Galician–German
3. *Iván Arias-Arias, María José Domínguez Vázquez, María Teresa Sanmarco Bande & Carlos Valcárcel Riveiro*
   From Content to Code and Back: Evaluating Agentic Coding in LLMs for Building R Shiny Dictionary Interfaces
4. *Petra Bago & Ivana Eterović*
   Compiling a 16th-Century Croatian Dictionary in the Age of AI: Challenges and Perspectives
5. *Martina Bajčić & Ana Ostroški Anić*
   Defining Legal Concepts in the AI Age: An Exploratory Study into Influencer Practices
6. *Elena Berthemet*
   Designing an Idiomatic Dictionary
7. *Harald Bichlmeier*
   Vol. IX of the *Etymological Dictionary of Old High German* [*Etymologisches Wörterbuch des Althochdeutschen (EWA)*]: History Structure, Contents
8. *Riccardo Billero*
   Lexicography in the Age of AI: Training Systems with Human Expertise in Art Terminology
9. *Ben Black Bear, Jr., Richard Two Dogs & Jan Ullrich*
   Building the Largest Dictionary for an Indigenous Language of North America
10. *Marek Blahuš, Miloš Jakubíček, Katarína Petreková & Yingtian Wu*
    Automatic Drafting of Terminological Dictionaries with One-Click Dictionary for 34 Languages
11. *Igor Boguslavsky, Vyacheslav Dikonov, Evgeniya Inshakova, Alexandre Lazursky, Svetlana Timoshenko & Tatiana Frolova*
    Lexicographic Resources for Inference-Based Semantic Text Processing System
12. *Ivana Brač*
    Semantic Class Assignment in a Croatian Verb Lexicon: Challenges of Applying VerbNet's Classes
13. *Katja Brankačkec & Barbora Štěpánková*
    Multifunctional Function Words in the Upper Sorbian-Czech Dictionary Mudra 2.0
14. *Cormac Breathnach & Pádraig Ó Mianáin*
    Reversing the New English–Irish Dictionary Database to Create Entry Frameworks for Irish–English and Monolingual Irish Dictionaries
15. *Rukiye Smilla Burkart & Susanne Kabatnik*
    Wie definiert man das Unbekannte? Prompt Engineering als Methode der KI-gestützten Bedeutungsdefinition im Pandemictionary
16. *Valeria Caruso & Geraint Paul Rees*
    From Lexical Theory to AI Translation: Procomplement Verbs and ChatGPT
17. *Lian Chen, Damien Nouvel, Huy-Linh Dao & Alexander Delaporte*
    From Corpus to Dictionary: OntoLex-Lemon-based modelling of neology in Chinese–French and Vietnamese–French computational lexicography

18. *Luisa Cimander*
English Adjectives in German: Processes of Integration since the 17th Century and Lexicographic Implications

19. *Li Cui, Seol Namkung, Jun Lee, Hae-Yun Jung & Kilim Nam*
What LLMs Learn and What They Miss: Frequency Effects in the Automated Description of Korean Verbal Inflection

20. *Flávia de Oliveira Maia-Pires & Silvio Licurgo Pires*
User-Oriented Specialized Digital Lexicography and Terminological Accessibility: The LexiC platform applied to the glossary of Brazilian migration legislation

21. *Gilles-Maurice de Schryver*
Chatbot Lexicography: Are GenAI LLMs simply 'dictionaries' in their own right and 'lexicographers' plain and simple?

22. *Gilles-Maurice de Schryver*
The Ninth Art in Lexicography: The power of comic strips

23. *Ida Dringó-Horváth & Katalin P. Márkus*
Dictionary Use in a Technology-Driven Era: Longitudinal Evidence from EFL and GFL Graduates

24. *Sophia Fiedler*
Towards an Interactional Lexicon of Spoken German: A Functional, Audio-Based Lexicographic Resource

25. *Ivana Filipović Petrović & Slobodan Beliga*
Modelling Phraseological Semantic Structures in DMLex:Aligning a Croatian Lexonomy-Based Idiom Dictionary with a Thematic Index

26. *Hanna Fischer, Alfred Lameli, Nathalie Mederake & Nico Urbach*
Linking Dictionaries Through Meaning: A sense-level evaluation of taxonomy-guided LLM classification in lexicography

27. *Margarita Freixas, Esther Hernández & Marija Žarković Eriksson*
e-*DRAE* 1884: Towards a Digital and Hypertextual Critical of a Nineteenth-Century Dictionary

28. *Christine Ganslmayer & Martina Nied Curcio*
How German Language Textbooks Address the Digital Transformation of Lexicography

29. *Marta Garcia Casado*
New Senses, New Methods: Clustering as a guide for detecting semantic neologisms

30. *Pol Garriga Martínez*
The VOcabulary of Mental Health: Exploitation of specialized diachronic corpora with artificial intelligence for qualitative semantic analysis

31. *Zoe Gavriilidou*
Spanish Loanwords in Modern Greek: Linguistic Analysis and Lexicographic Treatment in the Dictionary of Standard Modern Greek

32. *Zoe Gavriilidou & Evanthia Konstantinidou*
Pedagogical Lexicography in Practice: Raising Dictionary Use Strategies through Card-Based Instruction

33. *Alexander Geyken & Gregor Middell*
Trend Word Detection for DWDS, a Lexical Information System for Contemporary German

34. *Rufus H. Gouws & Theo J.D. Bothma*
Optimising Interface Design in Online Lexicographic Products

35. *Barbara Heinisch*
Dictionaries, Glossaries and Terminology Databases in the Age of Artificial Intelligence: Rethinking boundaries and functions

SHORT ABSTRACTS

36. *Ondřej Herman*
Sense Classification for Word Sketches

37. *Polina Iaroshenko & Natalia Loukachevitch*
Morphological Forms in Emotion Lexicons: Evidence from LLMs

38. *Ellert Thor Johannsson, Jóhannes B. Sigtryggsson & Thordis Ulfarsdottir*
Unlocking a Nineteenth Century Dictionary: A Closer Look at Konráð Gíslason's *Dönsk orðabók* (1851)

39. *Boris Kern*
Word-Formation Chains in Slovenian: Evaluating the Effectiveness of AI

40. *Václava Kettnerová, Veronika Kolářová & Jiří Mírovský*
Where the Boundary Between Derivational and Inflectional Morphology Meets Lexicography: Building the DeriVallex Lexicon

41. *Agnes Kim*
Digital Metalexicography as a Tool for Historical (Contact) Linguistics

42. *Veronika Kolářová, Václava Kettnerová, Michal Olbrich & Jiří Mírovský*
NomVallex: Valency of Czech nouns and adjectives in networks of derivationally related words

43. *Mariia Kopylova*
AI-Assisted Detection and Mitigation of Lexicographic Bias: Semantic Reduction in Definitions of Five Socio-Legal Terms

44. *Katharina Korecky-Kröll, Philipp Stöckle, Daniel Elsner, Wolfgang Koppensteiner & Elisabeth Eder*
LLM-Assisted Lexicography in Low-Resource Contexts: From prompt to dictionary article

45. *Iztok Kosem, Polona Gantar, Simon Krek & Tjaša Arčon*
Divide and Define with LLMs: Generating Definitions for the Digital Dictionary Database of Slovene

46. *František Kovařík, Marek Blahuš, Miloš Jakubíček & Vojtěch Kovář*
Agree to Disagree: Intra-annotator Agreement in Semi-Automatic Vocabulary Building

47. *Dominika Kováříková & Laura A. Janda*
Emotions in Context: A Semantic Field Viewed through Grammar and Collocations

48. *Barbara Lewandowska-Tomaszczyk*
The Integration of British and American Slang in Contemporary Polish and the Role of LLM-Driven Dynamic Lexicography

49. *David Lindemann*
Developing a Digital Version of Gidor Bilbao's Latin-Basque Dictionary

50. *Ivana Lovrić Jović & Lana Hudeček*
Synonyms That Never Met: On the Compilation of the Dictionary of the Dubrovnik Idiom

51. *Tinatin Margalitadze, Giorgi Meladze, Ketevan Mchedlishvili, Tamar Laluashvili & Giorgi Okropiridze*
Effectiveness of AI for the Lexicography of Low-Resource Languages: The Dictionary of Georgian Neologisms

52. *Aleksandra Marković, Cvetana Krstev, Milica Dinić Marinković & Ranka Stanković*
Towards a Repository of Similes in Contemporary Serbian

53. *Peter Meyer*
An AI-Driven Natural Language Interface for Graph Database Queries in a Lexicographical Resource

54. *Ana Mihaljević & Josip Mihaljević*
Artificial Intelligence as Support for the Digitization and Compilation of the *Dictionary of the Croatian Redaction of Church Slavonic*

55. *Tibor M. Pintér, Manana Rusieshvili-Cartledge, Dóra Pődör, Katalin P. Márkus & Marine Makhatadze*
Sublexical Systems of Colour in Biblical Georgian and Hungarian: Findings of the Pentateuch

56. *Bruno Nahod & Marijana Tuta Dujmović*
From Text to Terms: TermScout, an AI Module for Early Terminology Processing

57. *Gunta Nešpore-Bērzkalne, Mikus Grasmanis & Madara Stāde*
Possibilities of Representation of Multi-Word Expressions in the Latvian Electronic Dictionary *Tēzaurs*

58. *Sanni Nimb, Ida Flörke, Agnes Aggergaard Mikkelsen, Nathalie Hau Norman & Nicolai Hartvig Sørensen*
AI Used to Detect Problematic Content in Corpus Examples

59. *Krzysztof Nowak, Iwona Krawczyk & Jagoda Marszałek*
DiaLexPol: LLM-assisted Sense Assignment for a Diachronic Dictionary of Latin

60. *Krzysztof Nowak & Dorota Mika*
PoliLex: Modelling Heterogeneous Lexical Resources: Lessons from the Dictionary of Polish Dialects

61. *Emma Josefin Ölander Aadland*
A Critical Lexicographical Discourse Analysis of New Adjectival Entries in the Standard Dictionaries for Norwegian

62. *Benedikt Perak, Dragana Špica & Irena Srdanović*
A Graph-Based Lexicographic Workflow for Japanese Adjectives and Their LExical Relations in an AI Tutor

63. *Linda Prossliner*
AI-Supported Phraseography for Young Learners: Designing Definitions and Illustrations through AI – Opportunities and Challenges

64. *Laura Rebosio*
Ostensive Aids in Learner's Dictionaries: A diachronic, comparative analysis

65. *Lydia Risberg, Kristina Koppel, Margit Langemets, Hanna Maask, Esta Prangel, Maria Tuulik & Silver Vapper*
Are Bears More Offensive Than Dogs When Drunk? Using LLMs to systematize the labelling of Estonian synonyms

66. *Ines Röhrer & Manuel Raaf*
Reflections on the Semantic Understanding of Generative AI: From LLM experiments to lexicographical tools

67. *Antonio San Martín & Catherine Trekker*
Introducing Human-Centeredness in AI-Assisted Lexicography

68. *Saghar Sharifi & Abolfazl Alamdar*
Towards a Handbook of Persian Lexicography: A Lexicographical-Functions Analysis of 20th–21st Century Dictionaries

69. *Jóhannes B. Sigtryggsson*
The Effects of Purism on a Nineteenth-Century Icelandic Bilingual Dictionary

70. *Hana Skoumalová*
Expanding the MWE Database with Variants

71. *Deepak Solanki*
User-Oriented Specialized Lexicography: Corpus-Driven Models for Targeted Dictionary Design

72. *Michael Spagnol, John J. Camilleri & Dwayne Ellul*
From *Ḋamma* to *Ġabra*: Four Centuries of Maltese Lexicography

SHORT ABSTRACTS

73. *Vasyl Starko & Andriy Rysin*
Large-Scale Semiautomatic DEtection of New and Unregistered WOrds in Modern Ukrainian

74. *Nika Stefan, Johan van der Zwaag & Hindrik Sijens*
Beyond the Single Standard: Addressing Variation and User Needs in an Online Frisian Dictionary

75. *Petra Storjohann*
The Interplay Between Individual New Lexical Items and Multi-Layered Discourse Representations of Neologisms

76. *Silga Sviķe*
A Lexicographical Analysis of Neologism Dictionaries Based on the Multilingual Dictionary of New Words (2010–2024)

77. *Pius ten Hacken*
The Theoretical Foundation of Dictionary Definitions

78. *Dušica Terzić, Dejana Mirković-Birtašić & Saša Marjanović*
Towards a French–Serbian Dictionary of Sports Terms: Corpus-based or AI-Assisted Methodology?

79. *Carole Tiberius, Martin Kroon, Lut Colman & Dalí Chirino*
Ranking Collocations in *Woordcombinaties*

80. *Jesús Torres del Rey & M.ª Teresa Fuentes Morán*
A Hybrid Approach to Web Accessibility Evaluation in Digital Lexicography: Comparing and Integrating Rule-Based Validators and Generative AI

81. *Ágoston Tóth*
(UG)TPEX: A (Usage-Guided) Typical Example Score

82. *Ágoston Tóth & Esra Abdelzaher*
GPT as a Sense-Tagger for Selecting Dictionary Examples

83. *Lars Trap-Jensen & Henrik Lorentzen*
Revising the Usage-Labelling System of the Danish Dictionary

84. *Thordis Ulfarsdottir & Ellert Thor Johannsson*
The Reversal of a Bilingual Dictionary Using AI

85. *Anna Vacalopoulou*
A Definition Walks into a Dictionary: The lexicography of humor

86. *Ene Vainik, Heete Sahkai, Ahto Kiil, Jelena Kallas & Geda Paulsen*
Semantic Clustering and Categorisation of Constructional Collexemes Using LLMs

87. *Carlos Valcárcel Riveiro, María Teresa Sanmarco Bande, Iván Arias-Arias & María José Domínguez Vázquez*
Can Lexicographic Entries Embed Didactic Modules? A Case Study on LLM-Generated Outputs for a Dictionary Interface

88. *Martina Vokáčová & Anna Marklová*
Affordance Norms in Czech and English: Human–LLM Misalignment in Embodied Lexical Knowledge

89. *Zongyuan Wang, Valeria Caruso & Anna De Meo*
Collaborative Lexicography for Heritage Language Learning

90. *Tarrin Wills & Ellert Thor Johannsson*
Can AI-ASsisted Normalisation Improve Access to Historical Lexicographic Data?

91. *Shigeru Yamada*
Pedagogical Value and Insights from Comparing EFL Dictionaries: Evidence from *apparently*, *be on speaking terms*, and *suffer* (*from*)

92. *Yelena Yerznkyan & Diana Movsisyan*
    Encoding Emotivity in Dictionary Definitions: A Corpus-Based Study of Verbs of Understanding
93. *Marija Žarković Eriksson*
    Legal Lexicon and the e-*DRAE* 1884: A Diachronic Study of Specialised Vocabulary in a General Dictionary Using a Digital Critical Edition

Workshops:

- *Ilan Kernerman & Kris Heylen*
  8th Globalex Workshop on Lexicography and Neology
- *Voula Giouli, Athina Sioupi, Nina Böbel & Alexander Ziem*
  Constructicography in Practice: Constructions, Challenges, and Opportunities
- *Ondřej Matuška*
  Postediting Lexicography with Sketch Engine and Lexonomy
- *Lian Chen & Besim Kabashi*
  Multi-Word Expressions and Phraseology: Corpus-based and computer-processed

PART II

KEYNOTES

EURALEX 2026

Robert Lew

WITH BOTS LIKE THESE, WHO NEEDS DICTIONARIES?

Keywords Dictionaries vs. AI; generative AI; bilingual dictionaries; language learning; tool effectiveness; future of language resources

Four years after the public launch of ChatGPT, Generative AI has entered many domains, both private and professional, and it is not going away. As anticipated in a prescient presentation at eLex 2019 (The Sintra variations), AI bots have begun to compete with dictionaries as reference tools for language-related tasks. Yet despite the pervasiveness of this emerging technology, dictionaries remain important consultation tools for certain types of language problems, at least for advanced learners, who appear able to choose critically between dictionaries and AI tools depending on the task at hand. At the same time, recent studies comparing the effectiveness of chatbots and dictionaries suggest that AI chatbots are not necessarily more effective than dictionaries—especially bilingual dictionaries—whether measured in terms of immediate success or longer-term learning. Against this empirical backdrop, the talk will sketch the current and near-future roles of dictionaries and AI tools.

Contact information

Robert Lew
Adam Mickiewicz University, Poznań
rlew@amu.edu.pl

Kate Wild

BEYOND THE ENTRY

The Evolution of Historical Lexicography at the Oxford English Dictionary

Keywords Historical lexicography; Oxford English Dictionary; computational lexicography; AI-assisted lexicography; dictionary revision strategies; historical corpus development

As the Oxford English Dictionary approaches the centenary of the completion of its first edition, this talk reflects on how the practice and possibilities of historical lexicography have changed over the past century. The first edition of the OED was produced letter by letter and entry by entry, culminating in the publication of its final volume in 1928. Alphabetical order is no longer a constraint, and for the past two decades entries have been selected for revision according to criteria such as cultural significance, semantic complexity, and user feedback, enabling more dynamic and targeted updates. In recent years there has been a further development in that the individual entry is no longer the only, or even the primary, unit of revision. Alongside traditional entry-by-entry revision, the OED is now undertaking systematic improvements to components across the text, sometimes supported by computational methods and AI. Key goals for the centenary include the addition of recent quotation evidence to all senses in current use, the modernization of outdated defining language across the dictionary, and the inclusion of etymologies and variant forms lists in all entries.

Looking beyond the text itself, the OED is also engaged in initiatives that position it increasingly as a digital dataset as well as a dictionary. These include the development of a historical corpus both as a tool to support editorial work and as a resource to complement OED data; and (in collaboration with the Historical Thesaurus team at the University of Glasgow) the further development and improvement of the Historical Thesaurus taxonomy, both as a structured dataset linked to the OED and as a tool which can support historical word sense disambiguation and other types of data analysis.

The argument advanced in this talk is not that historical lexicography should abandon its entry-by-entry craft, but that its future depends on combining that craft with methods designed for the dynamic, linked, data-rich digital environments in which dictionaries now live.

Contact information

Kate Wild
Oxford University Press
kate.wild@oup.com

Andreas Baumann

POLYSEMY

Operationalization, Emergence, Perception, and the Human in the Loop

Keywords Polysemy; corpus frequency; sense similarity; language acquisition; lexical sentiment; human-in-the-loop

In lexica of natural languages, polysemy is the rule rather than the exception. Understanding why it is so prevalent—given that the presence of multiple senses for a single form, from a purely semiotic point of view, constitutes ambiguity—and how it emerges is central to lexicology. Addressing this question, of course, requires proper operationalization of polysemy in the first place. A robust, straight forward, and commonly used way of measuring the polysemy of a word is simply counting the number of sense entries in a lexicographically curated resource. Such an approach, however, neglects the relative frequencies of a word's senses as well as how similar senses are to each other. By integrating lexicographic and corpus data, I will discuss frequency and similarity aware measures of polysemy, originating from quantitative ecology (Leinster & Cobbold, 2012), and how they relate to ratings of subjectively perceived polysemy, gathered through crowdsourcing efforts. I will then discuss the roles that frequency and acquisition play in the emergence of polysemy (Baumann & Hartmann, 2026). Finally, I will zoom in on one specific dimension of meaning: lexical sentiment. I show that variation in human sentiment annotations correlates well with how emotional polysemy is subjectively perceived, but that purely NLP based approaches in fact struggle with appropriately capturing subjective emotional ambiguity. I take this to stress the relevance of 'the human in the loop' also in lexical and lexicographic research.

References

Baumann, A., & Hartmann, S. (2026). The chicken and the egg: unraveling aspects of semantic change and how they relate to lexical acquisition. *Cognition*, 266(2026), 1–20. https://doi.org/10.1016/j.cognition.2025.106301

Leinster, T., & Cobbold, C. A. (2012). Measuring diversity: the importance of species similarity. *Ecology*, 93(3), 477–489. https://doi.org/10.1890/10-2402.1

Contact information

Andreas Baumann
University of Vienna
andreas.baumann@univie.ac.at

Ana Salgado

DICTIONARIES AS PIPELINES

Lexicographic Work at the Lisbon Academy of Sciences

Keywords Lexicographic pipeline design; editorial governance; pluricentric language variation; corpus-driven lexicography; TEI/XML; cartography of variation

This keynote reframes dictionary-making as the design of a living, governed pipeline: a system that turns linguistic evidence into publishable lexical knowledge while keeping editorial decisions traceable and revisable. Drawing on current work at the Lisbon Academy of Sciences across a portfolio of initiatives, it shows how this pipeline supports complementary goals: a continuously updated Dicionário da Língua Portuguesa (DLP), inclusive and participatory lexicographic practices (ACL+), a collaborative lexicographic platform mapping variation across the Community of Portuguese Language Countries (CPLP) via the Atlas Lexicográfico da Língua Portuguesa (ALLP), and the reuse of lexical data in infrastructures and language technologies (PORTULAN).

The pipeline starts by converting legacy materials into structured, machine-actionable data (such as XML) and by applying consistent normalisation and TEI-based annotation. Evidence is then continuously gathered from corpora, query logs, and community contributions, and assessed through a combination of quantitative signals and expert editorial judgement. A central theme is governance: how inclusion and revision decisions are documented, reviewed, and made auditable—especially when balancing pluricentric Portuguese.

Extending the pipeline from dictionary entries to atlas-style knowledge, ALLP illustrates how lexical description can also support cartography of variation, combining curated entries with usage evidence and, where relevant, multimedia documentation. The keynote concludes with practical lessons—what scales well, what slows teams down, and how to monitor quality over time.

Contact information

Ana Salgado
University of Porto
acsalgado@letras.up.pt

Jan Hajič

LARGE LANGUAGE MODELS, SYMBOLIC MEANING REPRESENTATIONS AND ONTOLOGIES

Perspectives and Questions

Keywords Large Language Models; European AI initiatives; OpenEuroLLM; symbolic meaning representation; Uniform Meaning Representation; Prague Dependency Treebank

In the talk, I will present the basics of Large Language Models (LLMs) and the status of the European LLM developments, in particular the OpenEuroLLM project that aims to develop open, multilingual foundational models in Europe and its challenges. I will then move on to summarise some recent developments on the other side of the research spectrum: symbolic meaning representations and latest developments of the most prominent ones, such as the Uniform Meaning Representation and the Prague Dependency Treebank. Such representations make heavy use of lexical resources – and I will show how these can be linked together and unified for making current and future meaning representation annotation efforts more consistent, while preserving all the information and alignment between form (plain text) and the representations. Towards the end of the talk, I will connect the current work on LLMs and meaning representations by postulating new research questions for which we now have resources and tools to work on.

Contact information

Jan Hajič
Charles University Prague
hajic@ufal.mff.cuni.cz

PART III

EXTENDED ABSTRACTS

EURALEX 2026

Vassiliki Afentoulidou, Ioannis Tziogkas, Christos Papanagiotou, Aretousa Giannakou & Anastasia Christofidou

AI-ASSISTED LEXICOGRAPHY IN MODERN GREEK

Neology and Genre

Keywords AI-generated annotations; corpus-based lexicography; journalistic genre classification; machine learning; Modern Greek neology; monitor corpus; neology documentation

This contribution presents the main objectives and methodological commitments adopted at our Research Institution concerning Modern Greek neology and addresses the automatic classification of texts containing neologisms according to journalistic genre with the aid of AI-powered applications.

The research pipeline is organized around NEOΔHMIA (NEODEMIA), a digital environment comprising four constantly updated components: neologism text retrieval component, neologism extraction component, neologism classification component, and neologism monitoring component (cf. European Network on Lexical Innovation deliverables for standard practices). The *Monitor Corpus of Neologisms* (MCN), a specifically designed corpus of online newspaper discourse, nowadays comprising ~900 million words, is compiled following the TEI standards (Web for Corpus methodology).

Candidate neologisms are detected using computational methods based on stopword lists (including named-entity exclusion lists) and are refined via a manual selection process guided by lexicographic criteria, such as their (non)presence in Modern Greek reference dictionaries. Alongside lexical usage, morphological and textual metadata (genre, text type, topic, text structure, etc.) are documented, as well as webometric data capturing the spread of neologisms across domains and sources (Web as Corpus methodology). Validated items are then stored in a database and periodically published in the *Bulletin of Scientific Terminology and Neologisms* (for instance, see Christofidou, 2025).

One of the central issues of the research presented above concerns the multi-level interaction of neologisms with the topic, type and genre of the text in which they occur, where each shapes the interpretation and function of the other (Kerremans, 2015; Christofidou, 2018; Christofidou et al., in print). Such a discourse-based approach to neology could have important lexicological/semantic and thus lexicographical implications.

Specifically, the paper focuses on the neologism classification component, which hosts the global classification service of NEOΔHMIA (structured around metadata annotation schemes, such as genre, text type, topic, and URL source classification), and presents the implementation of a news-article genre classifier. The classifier relies on a structured 20-label genre taxonomy (e.g., news articles, book/product reviews, political interviews, recipes), developed for neologism documentation. When recorded in the database of neologisms, the annotated texts serve as neologism citations at the document level and dynamically feed into headword-level genre profiling of the corresponding neologisms.

Methodologically, two complementary strategies are adopted: (a) a fine-tuned GreekBERT model trained on approximately 5,000 annotated texts from MCN (19 predefined genre categories plus 1 open category) and (b) a generative AI approach based on GPT-x models and open-weight LLMs. We report the experimental procedure for both methods, namely data

preprocessing and text representation heuristics via deep contextualized embeddings, model selection, and prompt engineering (system and user prompts), as well as the evaluation metrics used. Pilot experiments already conducted on four main journalistic genres (news/informational articles, opinion articles, current affairs reports, humorous/observational columns) yielded promising results (Macro-F1: 82.25% *vs* 82.0% respectively; Afentoulidou et al., 2025).

We further discuss methodological challenges and opportunities in AI-assisted neology documentation, especially considering NEOΔHMIA's long-term goal of producing and hosting a *Digital and Corpus-based Dictionary of Greek Neologisms* (Afentoulidou & Christofidou, 2021). Within this lexicographic framework (de Schryver, 2023), corpus evidence and genre-based classification constitute core structural components of future entries. A model entry of such a dictionary of neologisms is presented, by highlighting how contextual annotations by AI systems can be incorporated in the lexicographic workflow.

References

Afentoulidou, V. & Christofidou, A. (2021). It's a long way to a dictionary: Towards a corpus-based dictionary of neologisms. In Z. Gavriilidou, L. Mitits & S. Kiosses (Eds.), *Lexicography for inclusion: Proceedings of the 19th EURALEX International Congress, 7–9 September 2021, Alexandroupolis* (Vol. 2, pp. 597–606). Democritus University of Thrace.

Afentoulidou, V., Tziogkas, I., Varlamis, I., Giannakou, A. & Christofidou, A. (2025). *Neologisms of Modern Greek: Monitor Corpus and genre classification*. Conference presentation at ICGL 17, Cambridge, September 2025.

Christofidou, A. (2018). Κειμενικά είδη και τρόποι λόγου: Μία συνολική κειμενογλωσσολογική πρόταση [Text genres and discourse modes: An integrated text-linguistic proposal]. In K. Ntinas (Ed.), *Figura in praesentia* (pp. 604–624). Patakis.

Christofidou, A. (Ed.) (2025). *Bulletin of Scientific Terminology and Neologisms* (Vol. 16–17). Academy of Athens. https://doi.org/10.5281/zenodo.15731933

Christofidou, A., Gkolfopoulos, G., Foufi, V., Afentoulidou, V., Giannakou, A. & Varlamis, I. (in print). Modern Greek neologisms: Monitor Corpus and automatic text classification. In *Proceedings of ICGL 16*. Aristotle University of Thessaloniki.

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.

European Network on Lexical Innovation (ENEOLI). COST Action Network. *Resources & Deliverables*. Retrieved May 25, 2026, from https://eneoli.eu/outcomes/

Kerremans, D. (2015). *A web of new words: A corpus-based study of the conventionalization process of English neologisms*. Peter Lang.

Contact information

Vassiliki Afentoulidou
National and Kapodistrian University of Athens & Academy of Athens
vafentoul@phil.uoa.gr

Ioannis Tziogkas
Harokopio University of Athens
john.tziog@gmail.com

Christos Papanagiotou
Academy of Athens
cpapanagiotou@academyofathens.gr

Aretousa Giannakou
Academy of Athens
giannakou.aretousa@ac.eap.gr

Anastasia Christofidou
National and Kapodistrian University of Athens & Academy of Athens
christo@academyofathens.gr

Mikel Alonso

AI-ASSISTED INFORMATION EXTRACTION FROM LARRAMENDI'S TRILINGUAL DICTIONARY (1745)

Keywords Legacy dictionaries; information extraction; TEI-Lex0; Basque; trilingual

This poster presents the research on the application of different approaches to the information extraction process and encoding of the 18th century Larramendi's Trilingual Dictionary. The digitisation and computational analysis of historical dictionaries present significant challenges due to archaic typefaces, complex layouts, and inconsistent lexicographic structures. Manuel de Larramendi's *Diccionario trilingüe castellano, bascuence y latin* (1745, hereafter LAR), a foundational text for modern Basque lexicography, remained largely inaccessible in a structured digital format, available only as scanned facsimile images.

Initial digitisation efforts employed a semi-automatic workflow combining machine learning (ML) for Optical Character Recognition (OCR) using *Kraken*[1] and rule-based information extraction. The OCR for LAR was completed, and the resulting transcription was uploaded to Wikisource[2], a crowdsourcing platform that provides an interface for collaborative facsimile transcriptions. That process is ongoing and will very soon lead to an error-free digital version of the original text. Wikisource transcription quality is ensured by its rigorous, community-driven validation process, as completed works usually go through a two-step human review. The rule-based information extraction process facilitated the identification of Spanish headword candidates using positional data (line indentation) and orthographic features (capitalisation corresponding to alphabet sections). The extracted data was further enriched by linking Spanish lemmata to a contemporary reference, the *Diccionario de Autoridades* (1726-1739), and mapping historical Basque forms to modern standard lexicons. This work demonstrated the feasibility of converting a historical printed dictionary into structured datasets and linked data, though it highlighted limitations in fully capturing the complex microstructure of entries through rule-based methods alone (Lindemann & Alonso, 2021).

The first approach to information extraction from LAR involved a hybrid methodology. For a more nuanced structural analysis, it was then employed the *Elexifier*[3] toolchain, a ML-based application. A sample of dictionary columns was manually annotated with tags from the TEI-Lex0 schema and used for training a model to predict and segment lexicographic components—such as headwords, Basque and Latin equivalents—across the entire dictionary text. This ML approach proved complementary to the rule-based method, showing high precision for Spanish headwords and Basque translations but lower accuracy for Latin equivalents and cross-references, revealing the complexity of the dictionary's microstructure.

Advancing beyond specialised ML tools like *Elexifier*, contemporary research explores the application of general-purpose Large Language Models (LLMs) with vision capabilities for similar tasks. Inspired by methodologies applied to Estonian-German dictionaries (Jürviste & Jakobson, 2025) or to Swedish historical patent cards using the GPT-4o model (Xie et al., 2025),

1 See http://kraken.re for further information about this OCR system optimized for historical and non-Latin script material.

2 Accessible at https://eu.wikisource.org/wiki/Hiztegi_Hirukoitza

3 See https://github.com/elexis-eu/elexifier#elexifier

a third approach involves using DeepSeek[4], an advanced LLM, for the information extraction and encoding of LAR. The extraction of key microstructural elements like headwords (Spanish), translation equivalents (Basque, Latin), part-of-speech indicators, and example phrases and sentences is demanded from standard files, such as XML or JSON, or from the wikitext format text from Wikisource, with the few-shot prompting strategy, which basically consists of showing the model a few examples of inputs and desired outputs before asking it to solve the task. A fourth approach involves employing Microsoft Copilot[5], another multimodal LLM, for the same extraction task. The process consists of feeding Copilot with the same input and few-shot prompting, requesting structured TEI-Lex0 output that represents LAR's tripartite entries.

Following an experimental framework used for historical dictionaries (Jürviste & Jakobson, 2025), this allows for an evaluation of different model performances on identical material. Key evaluation metrics include the structural correctness of the generated markup (adherence to TEI-Lex0) and the model's ability to correctly classify and relate the Spanish, Basque, and Latin lexical units within each entry. Preliminary analyses of the results suggest that DeepSeek encodes TEI-Lex0 schema with almost no syntax errors and can correct the dictionary's microstructural inconsistencies in the encoded output file, whereas Microsoft Copilot tends to fail to follow consistently the encoding schema showed in the given examples.

This comparative analysis extends the results of the mentioned studies about the strengths and weaknesses of different LLMs in handling the specific challenges posed by early modern bilingual/multilingual dictionary digitisation, particularly for languages with less digital representation. Nevertheless, it would be interesting for further digitisation processes to conduct experiments providing the tools with scanned images, because unlike the earlier workflow requiring separate OCR and segmentation stages, a vision-enabled LLM could potentially perform text recognition and semantic structuring in a single, integrated step. This approach promises a more direct path from facsimile to structured data, potentially reducing pipeline complexity and adapting more flexibly to layout irregularities.

In conclusion, the initial work on LAR established a viable pipeline and valuable datasets from rule-based and specialised ML tools, although the evolution to general-purpose (vision-enabled) LLMs represents a significant shift in historical dictionary digitisation methodology. Approaches leveraging LLMs aim to integrate OCR, segmentation, and semantic encoding into a more efficient workflow. This research not only advances the digitisation of a key Basque lexicographic resource but also can contribute to computational historical lexicography by evaluating and refining modern AI-driven extraction techniques.

References

Jürviste, M., & Jakobson, J. (2025). Vision-Enabled LLMs in Historical Lexicography: Digitising and Enriching Estonian-German Dictionaries from the 17th and 18th Centuries. *arXiv*. https://doi.org/10.48550/arXiv.2510.07931

Larramendi, M. (1745). *Diccionario trilingüe castellano, bascuence y latin dedicado a la M.N. y M.L. provincia de Guipuzcoa.* Bartholomé Riesgo y Montero.

Lindemann, D., & Alonso, M. (2021). A workflow for historical dictionary digitisation: Larramendi's Trilingual Dictionary. 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. 598–614). Lexical Computing CZ, s.r.o. https://elex.link/elex2021/wp-content/uploads/2021/08/eLex_2021_39_pp598-614.pdf

4 See https://www.deepseek.com/en/

5 See https://copilot.microsoft.com/

Real Academia Española (Ed.) (2013). *Diccionario de autoridades: 1726 - 1739*. JdeJ Editores.

Xie, Y., Mela, M. L., & Tell, F. (2025). Multimodal LLM-assisted Information Extraction from Historical Documents: The Case of Swedish Patent Cards (1945-1975) and ChatGPT. *Digital Humanities in the Nordic and Baltic Countries Publications*, 7(2). https://doi.org/10.5617/dhnbpub.12294

Contact information

Mikel Alonso
EHU – University of the Basque Country
mikel.alonsoa@ehu.eus

Maria Ammari

PEDAGOGY AS THE CORE OF EDUCATION

A Domain-Analytic Route to a Corpus-Based Dictionary

Keywords Academic disciplines; academic writing; corpus design; domain analysis; education; EAP; needs analysis; pedagogy; phraseology; representativeness; specialised lexicography

Mastering academic language remains a central challenge across university disciplines, as learners are expected to handle increasingly complex topics, genres, and stance-driven patterns, often with higher stakes for those working in an additional language. In response, corpus linguistics has reshaped lexicography (Rees, 2021) by evidencing language use not only as single lexical units, e.g., academic wordlists (Coxhead, 2016), but also as recurrent phraseological combinations, where corpus methods have been especially transformative (Paquot, 2015). Projects such as the Louvain EAP Dictionary (LEAP) demonstrate that online, needs-driven academic writing aids can now be structured around discipline-specific phraseological evidence (Granger & Paquot, 2015).

Despite growing interest in discipline-sensitive EAP resources (Rees, 2021; Ang & Tan, 2018), Education remains completely unexplored as a target domain for discipline-specific EAP lexicography. Education was not formally recognised as an academic discipline within universities until the twentieth century (Wyse, 2020), and its lexicon is distinctive: high-stakes yet conceptually diffuse, multidisciplinary, institutionally embedded, and strongly value-laden, with terminology that is often contested, relabelled, and unevenly stabilised across subfields. The practical need is substantial: beyond researchers and postgraduate students writing about pedagogy in English as an additional language, the global English-language teaching industry produces a large population of trainee teachers (e.g., CELTA, TESOL programmes) who must master pedagogical terminology in English. Against the United Nations SDG 4 (Quality Education), accessible and precise pedagogical language becomes more acute still, particularly in multilingual academic communities (Adipat & Chotikapanich, 2022).

Methodologically, the poster presents a needs-driven corpus design (Nesi, 2015; Flowerdew, 2013) in which the target domain was not imposed from a pre-existing taxonomy but derived empirically from the community itself. Semi-structured interviews with twelve members of the Pedagogy in Education Studies (PES) community of practice were analysed using reflexive thematic analysis (Braun & Clarke, 2021), yielding fifteen inductive thematic clusters. These were consolidated into five macro-themes, pedagogy and assessment; inclusion and diversity; teachers; curriculum and technology; and PES as discipline, which served as the topical architecture of the corpus and as seed-term generators for web-based corpus construction via BootCat (Baroni & Bernardini, 2004). The resulting PES corpus was then evaluated retrospectively against Egbert, Biber and Gray's (2022) representativeness framework, following Kemp's (2024) concept of a Representativeness Argument: a structured post-hoc validation addressing both target-domain coverage and linguistic stability (Gray, Egbert & Biber, 2017). The poster visualises this full workflow, from community needs analysis through corpus construction to domain evaluation and the resulting pathway for extracting a pedagogy-centred lexicon that captures both high-frequency academic vocabulary and the phraseology through which pedagogical knowledge is routinely articulated by PES scholars.

Finally, the study is motivated by a commitment to strengthening Education's standing as an academic discipline: scholars who are driven by an ethical purpose to make a difference in the lives of learners and society constitute, in Furlong's (2013) terms, a hugely important force in

defining the field. Lexicographic resources that clarify pedagogy's core lexis and phraseology can function as enabling infrastructure for this purpose, supporting educational participation and knowledge exchange beyond institutionalised borders (Liguori et al., 2023). The poster will also address how the impact of such a dictionary might best be measured, including uptake metrics, user studies, and integration with existing EAP writing tools.

References

Adipat, S., & Chotikapanich, R. (2022). Sustainable development goal 4: An education goal to achieve equitable quality education. *Academic Journal of Interdisciplinary Studies*, 11(6), 174–183.

Ang, L., & Tan, K. (2018). Specificity in English for Academic Purposes (EAP): A corpus analysis of lexical bundles in academic writing. *3L: Language, Linguistics, Literature*, 24(2), 82–94.

Baroni, M., & Bernardini, S. (2004). BootCaT: Bootstrapping corpora and terms from the web. In M. T. Lino, M. F. Xavier, F. Ferreira, R. Costa & R. Silva (Eds.), *Proceedings of the Fourth International Conference on Language Resources and Evaluation* (pp. 1313–1316). ELRA.

Braun, V., & Clarke, V. (2021). *Thematic analysis: A practical guide*. SAGE.

Coxhead, A. (2016). Reflecting on Coxhead (2000), "A new academic word list". *TESOL Quarterly*, 50(1), 181–185.

Egbert, J., Biber, D., & Gray, B. (2022). *Designing and evaluating language corpora: A practical framework for corpus representativeness*. Cambridge University Press. https://doi.org/10.1017/9781316584880

Flowerdew, L. (2013). Needs analysis and curriculum development in ESP. In B. Paltridge & S. Starfield (Eds.), *The handbook of English for specific purposes* (pp. 325–346). Wiley-Blackwell.

Furlong, J. (2013). *Education – An anatomy of the discipline: Rescuing the university project*. Routledge. https://doi.org/10.4324/9780203078853

Granger, S., & Paquot, M. (2015). Electronic lexicography goes local: Design and structures of a needs-driven online academic writing aid. *Lexicographica*, 31(1), 118–141. https://doi.org/10.1515/lexi-2015-0007

Gray, B., Egbert, J., & Biber, D. (2017). Exploring methods for evaluating corpus representativeness. In *Corpus Linguistics International Conference 2017*. University of Birmingham.

Kemp, J. (2024). How do I know this Law corpus is reliable and valid? Using a representativeness argument for corpus validation. *Applied Corpus Linguistics*, 4(3), 100099. https://doi.org/10.1016/j.acorp.2024.100099

Liguori, A., Rappoport, P., Poce, A. & Rabagliati, M. (2023). Editorial: Towards 2030: sustainable development goal 4: quality education. A communication perspective. *Frontiers in Communication*, 8. https://doi.org/10.3389/fcomm.2023.1328002

Nesi, H. (2015). ESP corpus construction: A plea for a needs-driven approach. *ASp – la revue du GERAS*, 68, 7–23.

Paquot, M. (2015). Lexicography and phraseology. In D. Biber & R. Reppen (Eds.), *The Cambridge handbook of corpus linguistics* (pp. 460–477). Cambridge University Press.

Rees, G. P. (2021). Discipline-specific academic phraseology: Corpus evidence and potential applications. In M. Charles & A. Frankenberg-Garcia (Eds.), *Corpora in ESP/EAP Writing Instruction* (pp. 32–54). Routledge.

Wyse, D. (2020). Presidential address: The academic discipline of education. *British Educational Research Journal*, 46(1), 6–25.

Wyse, D., Brown, C., Oliver, S., & Pobleté, X. (2021). Education research and educational practice: The qualities of a close relationship. *British Educational Research Journal*, 47(6), 1466–1489.

PES corpus: *Pedagogy in Education Studies*. Retrieved from http://ske.li/pes_corpus

Contact information

Maria Ammari
Adam Mickiewicz University
maria.ammari@amu.edu.pl

Michela Bandini, Silvia Piccini, Andrea Bellandi & Emiliano Giovannetti

TOWARD A NARRATIVE-ORIENTED DICTIONARY

Computational Modelling of Youth Language in Socially Vulnerable Contexts

Keywords Narrative dictionary; youth deviance; Onto-Lex Lemon

This paper presents a computational narrative dictionary of youth language in contexts of high socio-economic vulnerability, where "narrative" is understood, following Bruner's distinction between paradigmatic and narrative modes of meaning-making (Bruner, 1986), as meaning grounded in experience, examples, values, and self-positioning rather than only in categorisation. The resource is designed for social and community workers, educators, and psychologists who work with young people in fragile settings and need to interpret not only what words denote, but also the experiences, values, and expectations that speakers attach to them. Its aim is therefore both descriptive and practical: to document situated meanings and make them usable in social and educational intervention. The dictionary does not replace standard lexicographical description; rather, it complements it with information that is usually excluded from dictionary entries but is central to intervention work: emotions, examples, social roles, expectations, and forms of self-positioning.

The dictionary is being developed within the project "Futuri (im)possibili: diagnosing the present and imagining the future" with the youth of Caivano[1], carried out in the metropolitan area of Naples, Italy. The project adopts an Action-Research approach (Elliot, 1993; Lewin 1946), combining empirical investigation with community-based social intervention. In line with user-oriented lexicography, the dictionary addresses concrete needs: it supports professionals in recognising semantic shifts and resemantisations, and in designing interventions that can help young people imagine alternative and more positive futures.

Building on Bruner's distinction, the resource focuses on socially sensitive Italian terms connected with power, deviance, gender, violence, and identity, such as 'aborto', 'criminale', 'femmina', 'infame', and 'soldi'. While standard dictionaries usually formalise meanings through paradigmatic definitions based on categorisation, synonymy, and semantic inclusion, the proposed dictionary also records associative networks, lived experience, and identity-related narratives. For instance, the word 'femmina' may activate biological or dictionary-like meanings, while also eliciting narratives of strength and care, but also inferiority, sexualisation, and violence. Meaning is thus treated as a cognitive and cultural construct shaped by shared representations, stereotypes, and lived experience.

The empirical basis consists of 58 semi-structured interviews (Adams, 2015), corresponding to 61 recordings and over 30,000 tokens of transcribed speech, collected mainly in street contexts in Caivano by outreach volunteers. Participants were aged 12 to 30, with a smaller group over 30, and were mostly residents of the area. Interviews were audio/video recorded, digitally transcribed, anonymised to ensure privacy protection, and organised in a structured corpus with

[1] Promoted by the Rut Foundation (https://fondazionerut.org/) in collaboration with the Don Calabria Foundation for Social Affairs (https://www.focsiv.it/fondazione-don-calabria-per-il-sociale-ets)

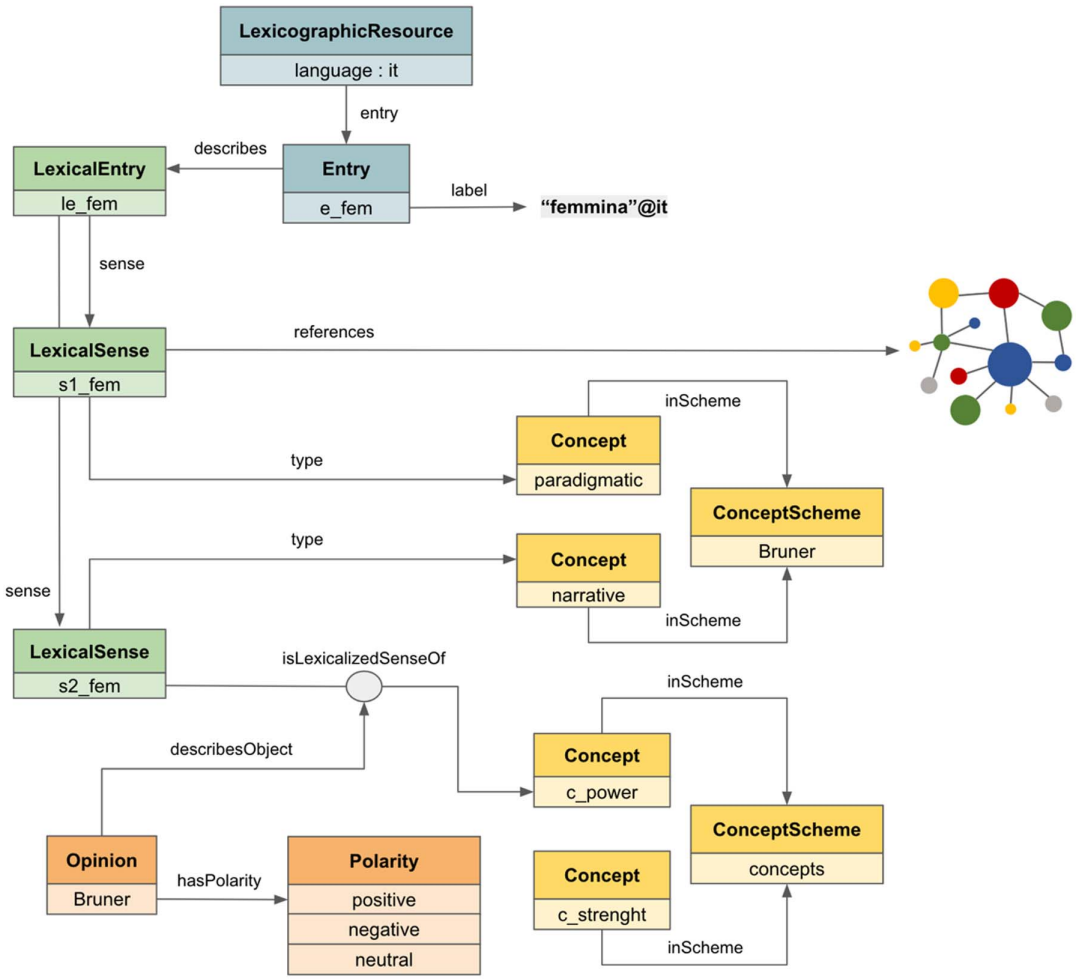


FIGURE 1 The proposed data model

metadata and unique identifiers. Automated transcription was combined with manual revision to guarantee accuracy. At the time of writing, the dictionary contains 16 lexical entries with both paradigmatic and narrative meanings.

The dictionary has been provided with a computational representation following OntoLex-Lemon[2], the *de facto* standard for lexical resources in the Semantic Web. This choice supports FAIR principles (Wilkinson et al., 2016) and facilitates interoperability with resources in the Linguistic Linked Data cloud. Each entry is modelled as a `lexicog:Entry` linked to an `ontolex:LexicalEntry`, while meanings are represented as `ontolex:LexicalSense`. Paradigmatic and narrative senses are encoded as SKOS concepts in a dedicated `skos:ConceptScheme` and assigned through `dc:type`. In the case of 'femmina' (Figure 1), the paradigmatic sense is linked to the concept WOMAN, whereas narrative senses are connected to concepts such as STRENGTH and POWER. Evaluative polarity is modelled through the MARL Opinion Ontology[3] by reifying the `ontolex:isLexicalizedSenseOf` relation.

The model is evaluated through competency questions that reflect expected use cases. These include, for example, retrieving concepts associated with narrative senses for a given entry,

[2] https://www.w3.org/2016/05/ontolex/
[3] https://www.gsi.upm.es/ontologies/marl/

listing concepts connected with negative or positive polarity, and identifying all narrative senses related to a specific concept. Rather than treating polarity as a simple label attached to an entry, the model relates it to specific senses and concepts, since the same word may carry different, even conflicting, evaluations depending on the speaker's narrative frame and social experience.

Future work will connect senses to textual occurrences in the corpus through OntoLex-FrAC[4] and enrich interview metadata using standard vocabularies such as Dublin Core, FOAF, and PROV-O. Integrating lexical, conceptual, textual, and metadata layers will enable cross-layer queries over the dataset, for example to examine which narrative senses and associations emerge in relation to speakers' age and gender. This can provide evidence-based insights for professionals working with youth in socially fragile contexts.

References

Adams, W. C. (2015). Conducting semi-structured interviews. In J. Wholey, H. Hatry & K. Newcomer (Eds.), *Handbook of Practical Program Evaluation* (pp. 492-505). Wiley.

Bruner, J. (1986). *Actual Minds, Possible Worlds*. Harvard University Press.

Elliott, J. (1993). *La ricerca azione.* Bollati Boringhieri.

Lewin, K. (1946). Action research and minority problems. *Journal of Social Issues*, 2(4), 34-46.

Wilkinson, M. D., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data,* 3.

Contact information

Michela Bandini
Istituto di Linguistica Computazionale "A. Zampolli"
michela.bandini@ilc.cnr.it

Silvia Piccini
Istituto di Linguistica Computazionale "A. Zampolli"
silvia.piccini@ilc.cnr.it

Andrea Bellandi
Istituto di Linguistica Computazionale "A. Zampolli"
andrea.bellandi@ilc.cnr.it

Emiliano Giovannetti
Istituto di Linguistica Computazionale "A. Zampolli"
emiliano.giovannetti@ilc.cnr.it

Acknowledgements

Scientific publication produced thanks to the agreement between the National Research Council – Institute of Computational Linguistics and the Rut Foundation.

[4] https://acoli-repo.github.io/ontolex-frac/

Biljana Božinovski, Makimi Kano, Peter Gobel & Mihaela Franjić

LEXICOGRAPHIC FUNCTIONS WITHOUT DICTIONARIES

GenAI-Mediated Lexical Problem-Solving Among Non-English-Major Tertiary EFL Students

Keywords Generative AI in EFL; lexical problem-solving; dictionary use vs. AI; EFL learners; AI-assisted language learning; invisible lexicography

With the increasing presence of generative artificial intelligence (GenAI) in learning and teaching practices, English as a Foreign Language (EFL) pedagogy is being fundamentally reshaped. In particular, GenAI has emerged as a core resource influencing students' approaches to learning and knowledge construction (Johnston et al., 2024; Ravšelj et al., 2025). Recent research indicates that students increasingly rely on GenAI chatbots not only for general writing support but also to resolve lexical problems and vocabulary challenges (Franjić & Božinovski, 2025; Xiao, 2024). Several studies have shown that large language model (LLM)-based chatbots such as ChatGPT can outperform traditional learner dictionaries in receptive and productive lexical tasks, especially in terms of speed and perceived usefulness (Božić Lenard & Šokčević, 2024; De Schryver, 2023; Lew et al., 2024; Ptasznik et al., 2024).

This is not to say that GenAI chatbots offer the same quality of lexicographic data as traditional dictionaries; rather, quality may no longer be a decisive factor for users anymore (de Schryver, 2023). This raises a critical question for lexicography: if dictionaries are no longer viable tools for most users, what role—if any—do lexicographic functions and knowledge still play in GenAI-mediated language use?

While a growing body of research has compared GenAI tools and dictionaries, it has largely focused on English-major students or controlled task-based comparisons (Chen, 2025; de Schryver, 2023; Lew et al., 2024; Mai, 2024; Ptasznik et al., 2024). Much less is known about non-English-major tertiary EFL students—who often lack dictionary literacy and increasingly rely on GenAI as their primary lexical access point. While some recent studies have begun to address GenAI use among ESP/EFL learners outside language-specialist programmes (e.g. Božić Lenard & Šokčević, 2024; Zhong, 2025), research remains limited.

This small-scale, user-centred study investigates lexicographic functions under GenAI mediation rather than dictionary use per se. Conducted in classroom settings in April 2026, it draws on data from approximately 40 non-English-major tertiary EFL students enrolled in English-language courses in Slovenia and Japan, all of whom regularly use GenAI tools and have received no formal dictionary training. The study examines (1) how students self-report their use of GenAI for lexical purposes, (2) how these perceptions align with observed lexical judgement when evaluating short GenAI-like texts containing collocational, patterning, and register-related issues, and (3) whether brief GenAI-mediated vocabulary exploration is associated with immediate uptake of target lexical items in a subsequent paper-based task.

The study adopts a one-session classroom design combining a brief survey (based on Ptasznik & Lew, 2025) with two task-based components (cf. Chen, 2025). First, a lexical judgement task completed without GenAI probes students' ability to identify and revise near-miss collocations, grammatical patterning errors, and register mismatches, allowing comparison between self-reported confidence and actual lexical evaluation behaviour. This is followed by a

time-limited GenAI-for-learning phase, in which students use a GenAI tool of their choice to explore selected lexical expressions, and a subsequent paper-based uptake task assessing their ability to use these items appropriately without tool support. The design enables triangulation of beliefs about GenAI as a lexical resource with observable judgement and short-term uptake, offering insights into the extent to which GenAI functions as a dictionary-like interface in L2 vocabulary learning and use.

The findings are expected to show that, even when dictionaries disappear as explicit tools in everyday writing practices, lexicographic functions remain central but become opaque under GenAI mediation ('invisible lexicography'). In contemporary EFL contexts, the pedagogical relevance of lexicography may lie in sensitising learners to the lexical dimensions of AI output—such as sense, collocation, patterning, and register—that require human interpretation and evaluation.

References

Božić Lenard, D. & Šokčević, K. (2024). ESP students' attitudes towards using digital dictionaries and AI-based tools in language learning. *e-mentor*, *107*(5).

Chen, Y. (2025). A Comparative Study on the Effectiveness of AI Chatbots and Dictionary Apps for Lexical Tasks and Retention. *Lexikos*, *35*, 157–182.

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.

Johnston, H., Wells, R. F., Shanks, E. M., Boey, T., & Parsons, B. N. (2024). Student perspectives on the use of generative artificial intelligence technologies in higher education. *International Journal for Educational Integrity*, *20*(1), 2.

Lew, R. (2024). Dictionaries and lexicography in the AI era. *Humanities and Social Sciences Communications*, *11*(1), 1-8.

Lew, R., Ptasznik, B., & Wolfer, S. (2024). The effectiveness of ChatGPT as a lexical tool for English, compared with a bilingual dictionary and a monolingual learner's dictionary. *Humanities and Social Sciences Communications*, *11*(1), 1-10.

Mai, H. V. (2024). Examining English-Majored Students' Performance and Attitudes towards ChatGPT and an Online Dictionary for Lookup. In V. P. H. Pham & J. White (Eds.), *Proceedings of the 21st AsiaCALL International Conference* (Vol. 6) (pp. 144-156). ICTE Press.

Ptasznik, B., Wolfer, S., & Lew, R. (2024). A learners' dictionary versus ChatGPT in receptive and productive lexical tasks. *International Journal of Lexicography*, *37*(3), 322-336.

Ravšelj, D., Keržič, D., Tomaževič, N., Umek, L., Brezovar, N., Iahad, N. A. & Aristovnik, A. (2025). Higher education students' perceptions of ChatGPT: A global study of early reactions. *PLOS ONE*, *20*(2), e0315011.

Xiao, Q. (2024, April). ChatGPT as an artificial intelligence (AI) writing assistant for EFL learners: An exploratory study of its effects on English writing proficiency. In *Proceedings of the 2024 9th International conference on Information and Education Innovations* (ICEIEI 2024) (pp. 51–56). Association for Computing Machinery.

Zhong, A. (2025). Prompts for Language Learners: A Practical Guide to Using DeepSeek as a Dictionary. *Lexikos*, *35*, 274-285.

Contact information

Biljana Božinovski
University of Maribor
biljana.bozinovski@um.si

Makimi Kano
Kyoto Sangyo University
makimik@cc.kyoto-su.ac.jp

Peter Gobel
Kyoto Sangyo University
pgobel@cc.kyoto-su.ac.jp

Mihaela Franjić
University of Maribor
mihaela.franjic1@um.si

Gabriela Brůhová & Veronika Raušová

THE "HELL OF A" CONSTRUCTION IN CASUAL WRITTEN ENGLISH ONLINE

A Synchronic Functional Account

Keywords Hell of a; functional classification; pattern-based taxonomy; procedural meaning; evaluation; intensification; quantification

Describing and classifying colloquial constructions that contribute to non-propositional meaning in discourse poses a persistent challenge for lexicography, particularly where interpretation depends on contextual inference rather than on stable form–meaning correspondences. Drawing on a 165-million-token corpus of Reddit posts and comments representing contemporary casual written English online, this paper explores the uses of the construction *a/one hell of a* and its orthographically reduced forms *helluva* and *hella* (cf. *hella*, OED, n.d.), with the aim of describing and operationalizing their synchronic functional range, building on ten Wolde's (2023) and Brůhová and Vašků's (2025) research into *of*-binominal phrases.

A grammaticalization cline of the construction proposed by ten Wolde (2023) distinguishes three recurrent uses of *a/one hell of a*. In the first use, the construction occurs in evaluative binominal noun phrases, where *hell* is interpreted literally or figuratively as attributing hellish, and by extension extremely bad, properties to the following noun (e.g. *This is a hell of a council of war*). In the second use, it functions as a broadened evaluative modifier (e.g. *They were having a hell of a time*), encoding a positive or negative extreme on a scale whose properties and interpretation are determined by the modified noun phrase and context (ten Wolde, 2023, p. 74). In the third use, the construction serves as a binominal intensifier modifying scalar adjectival premodifiers of nouns (e.g. *I've got a hell of a good bulldog*). The construction therefore operates with both implicitly and explicitly expressed scales "on which the speaker locates the degree of properties, which is a matter of speaker assessment and stance" (Ghesquière, 2014, p. 72). The process of grammaticalization of *a/one hell of a* therefore involves subjectification, with the bleaching of the descriptive meaning and the strengthening of the speaker's perspective (Jucker, 2010, p. 116).

Combining ten Wolde's (2023) diachronic account with the functional classification of binominal constructions proposed by Brůhová and Vašků (2025), and adapting both to the specific case of *a/one hell of a* and its reduced variants, we argue that given the semantic bleaching of *hell*, maintaining multiple evaluative distinctions or a separate descriptive function is neither necessary nor analytically productive for the analysis of contemporary usage. Instead, we distinguish three recurrent functions: evaluative, intensifying, and quantifying.

The consolidated evaluative function is attested for all three examined forms of the construction. Its defining property is the absence of an overt scalar or gradable element. In terms of encoded meaning, *hell* contributes only the notion of extremeness, while procedurally the construction prompts addressees to pragmatically infer a contextually appropriate evaluative scale on which this extreme value is to be interpreted, rather than encoding polarity or a specific evaluation independent of context. The category comprises all uses of *a/one hell of a* preceding nouns without modifiers, as well as nouns with non-gradable, non-scalar modifiers, in which it scopes over the NP as a whole (*This whole thread has been a hell of a rabbit hole . . .*). *Helluva* and *hella* perform the evaluative function as well, with *hella* only when preceded by an indefinite article (*And nostalgia is a helluva drug./. . .you made a hella profit*).

The second functional category represents intensifying use of *a/one hell of a* and its reduced forms. In this category, the construction intensifies an explicitly expressed scalar or gradable property, typically realized by an adjectival premodifier of nouns (*That's a hell of a long drive*), The intensifier *a/one hell of* functions as a booster, meaning "to a high degree" (ten Wolde, 2023, p. 109), and is felicitously paraphrasable by very. The construction frequently co-occurs with *a lot*, intensifying either quantity or degree, and is in these cases paraphrasable by quite, whether *a lot* functions as a pronominal quantifier (*And young men have a hell of a lot of problems . . .*) or as a degree adverb (*He seems a hell of a lot happier than you. / It helps a hell of a lot.*). The reduced forms *helluva* and *hella* display broader distribution beyond the noun phrase. Both can intensify adjectival phrases (*This is helluva impressive. / That's hella dope!*) and *hella* is additionally attested as an intensifier of adverbial and verb phrases (*. . . wake up hella early to make him breakfast . . . / I hella respect you for this response*).

Finally, *hella*, unlike the full construction *a/one hell of a* and *helluva*, performs a quantifying function. In this use, it occurs with both countable (*I had hella snacks waiting for me*) and uncountable nouns (*But yea they got hella charisma . . .*) and is paraphrasable as *a lot of*.

The paper proposes an operationalized, pattern-based classification of the functions performed by *a/one hell of a* and its reduced forms in contemporary casual written English online. This classification constitutes a first step toward implementing these distinctions as deterministic logic in an automated analysis pipeline, enabling systematic and replicable identification of functions.

References

Brůhová, G., & Vašků, K. (2025). English binominal noun phrases as seen through their Czech counterparts. *Linguistica Pragensia*, 35(2), 146–160. https://doi.org/10.14712/18059635.2025.2.4

Ghesquière, L. (2014). *The directionality of (inter)subjectification in the English noun phrase: Pathways of change.* De Gruyter. https://doi.org/10.1515/9783110338751

Jucker, A. H. (2010). Historical pragmatics. In M. Fried, J.-O. Östman, & J. Verschueren (Eds.), *Handbook of pragmatics highlights* (Vol. 6, pp. 110–122). John Benjamins Publishing Company. https://doi.org/10.1075/hoph.6.09juc

ten Wolde, E. (2023). *The English binominal noun phrase: A cognitive-functional approach.* Cambridge University Press. https://doi.org/10.1017/9781108921893

Contact information

Gabriela Brůhová
Charles University
gabriela.bruhova@ff.cuni.cz

Veronika Raušová
Charles Univeristy
veronika.rausova@ff.cuni.cz

Edmond Cane

LEARNING LIKE HUMANS, PARSING UNLIKE MACHINES

A Construction-Based Approach to Lexical Annotation

Keywords Construction Grammar; constructicon; lexical resources; linguistic annotation; information packaging

Background and aim

Large language models generate fluent, dictionary-like and analytically plausible text, yet their linguistic knowledge is encoded implicitly in numerical representations rather than as inspectable lexical entries, paradigms, feature structures or constructions. Classical annotation frameworks make categories explicit, but often reduce complex linguistic behaviour to flat tags and relations. This creates a representational gap for lexicography: lexical knowledge must be searchable, correctable, reusable and explainable, not merely recoverable from model output. Drawing on Construction Grammar (Goldberg, 2006; Croft, 2022), this paper presents an experimental model for Albanian that restores the lexicon to the beginning of processing, while redefining it as a constructicon: a structured inventory of lexical and grammatical form-content pairings. The aim is not to replace large language models or Universal Dependencies (UD; de Marneffe et al., 2021), whose purposes differ, but to develop a linguistically accountable architecture for richer lexical resources and annotation.

Model and method

The model represents constructions as transparent nested Python structures that can later be exchanged through JSON, XML or RDF-oriented formats. Each entry contains a compound_set with three components: constituent_set, which records the internal morphosyntactic parts; pos_set, which assigns a category to the complete entry or phrase; and feature_sets, which stores one or more compatible feature bundles. For nouns and noun-associated categories, the current feature inventory comprises gender, case, definiteness and number. A value may be concrete or left open as "ok", allowing another construction to supply it during unification. Two lexicons are presently distinguished: lexical-category dictionaries for nouns, adjectives, pronouns, prepositions, verbs and related classes, and marker dictionaries for inflectional elements. In a deliberate refinement of broader Construction Grammar terminology, form is restricted here to the graphic or phonological matter that activates an entry, whereas lexical-grammatical information is organized within content. Processing matches textual forms to stored constructions, tests compatibility, and builds larger usage-level constructs through licensed unification.

Illustration and preliminary evaluation

The Albanian noun *shef* 'chief' illustrates the procedure. The lemma is a lexical construction which contains information about (masculine) gender but leaves case, number and definiteness open. Unification with the postposed marker *-i* produces *shef-i* 'the chief', analysed internally as a noun plus marker and externally as a noun phrase. The surface form initially licenses

more than one feature analysis; in the larger construct *shef-i i zbulim-it* 'the chief of intelligence', contextual compatibility eliminates the unavailable reading, thus solving the ambiguity. The same mechanism models adjective and prepositional phrases, numerals, and the distinct processing roles of Albanian postposed and preceding articles. Grammatical agreement is thus represented as alignment information distributed across constructions, rather than as a label checked only after parsing. Initial domain-limited trials on nominal, adjectival, prepositional and selected verbal structures exceeded 90% accuracy and rose above 95% after missed constructions and compatibility patterns were corrected. These figures remain provisional because the test material, evaluation unit and error distribution have not yet been standardized.

Contribution and future development

The principal contribution is architectural. Annotation becomes the transparent result of recognizing stored constructions and projecting their information into constructs, rather than a separate layer of labels attached to bare tokens. This relocates POS and feature information within a richer representation and permits ambiguity without premature disambiguation. For Albanian lexicography, the approach supports a transition from digital dictionary to constructicon (Lyngfelt, 2018), encompassing lexical items, markers, phraseological patterns and larger constructions across the lexicon-grammar continuum. The present implementation is nevertheless incomplete: semantic frames, corpus-based frequency and productivity, broader verbal coverage, and external benchmark evaluation remain future tasks. Because its hierarchy is regular and explicit, the resource can also be mapped to linked-data models (Cimiano et al., 2016) and connected with corpus examples, ontologies, semantic frames and cross-linguistic correspondences. It therefore offers a practical basis for an interpretable, extensible and interoperable Albanian lexical resource.

References

Cimiano, P., McCrae, J. P., & Buitelaar, P. (2016). *Lexicon Model for Ontologies: Community Report*. W3C Ontology-Lexica Community Group.

Croft, W. (2022). *Morphosyntax: Constructions of the World's Languages*. Cambridge University Press.

de Marneffe, M.-C., Manning, C. D., Nivre, J., & Zeman, D. (2021). *Universal Dependencies*. Computational Linguistics, 47(2), 255-308.

Goldberg, A. E. (2006). *Constructions at Work: The Nature of Generalization in Language*. Oxford University Press.

Lyngfelt, B. (2018). Introduction: Constructicons and constructicography. In B. Lyngfelt, L. Borin, K. Ohara, & T. T. Torrent (Eds.), *Constructicography: Constructicon Development across Languages* (pp. 1-18). John Benjamins.

Contact information

Edmond Cane
Luarasi University
edmond.cane@luarasi-univ-edu.al

Elena Castaldi

INFLECTED-FORM SEARCH IN ANCIENT GREEK DICTIONARIES

A Form-To-Lemma Access Model

Keywords Ancient Greek; digital lexicography; inflected-form search; computational morphology; dictionary access

Dictionary consultation in Ancient Greek presupposes the ability to reconstruct canonical lemmas from highly inflected word forms, a task that can be challenging for users at any level. The lemma-based access model was a material constraint in printed dictionaries, but it can be addressed in digital dictionaries by exploiting the possibilities offered by computational morphology. The aim of this prototype is to link inflected Ancient Greek word forms to their corresponding dictionary lemmas, allowing users to search directly from inflected forms rather than exclusively from canonical lemmas. Although presented as a prototype, the model is designed to be reusable and applicable to any digital Ancient Greek dictionary.

Tools enabling morphological analysis of Ancient Greek forms already exist, most notably the Greek Word Study Tool of Perseus Digital Library[1] and the Morpho function on Logeion[2]. However, in both Logeion[3] and the Perseus Digital Library[4], dictionary consultation and morphological analysis remain functionally separate. In Logeion, the dictionary interface allows access only to forms attested in its database, while morphological analysis is provided through a separate tool (Morpho). Similarly, in the Perseus Digital Library, the Greek Word Study Tool provides automatic lemmatization and analysis, while dictionary resources such as the *Liddell-Scott-Jones Greek-English Lexicon*[5] are accessed independently, without systematic interaction between the two components. Although the output of morphological analysers typically provides a lemma linked to a dictionary entry, dictionary access and morphological analysis are presented through separate interfaces and rely on independent datasets: dictionary access is based exclusively on the headword inventory, while the morphological analyser operates independently of this inventory.

The model addresses this limitation by enabling dictionary access through inflected forms rather than exclusively through canonical lemmas. When an inflected word form is entered, the system first searches it within a curated dictionary inventory, which includes both canonical lemmas and selected inflected forms. If the input is found, the system directly returns the corresponding lemma. Only if no match is found is the form processed through an automatic morphological analyser.

At the core of the prototype lies a curated dataset linking inflected forms to dictionary-approved lemmas and their morphological analysis. In its current implementation, the dataset does not aim to represent a full dictionary inventory, but focuses on Ancient Greek polystemic (multi-stem) verbs, a particularly challenging category due to stem alternation and irregular morphology. This restricted scope allows the prototype to address a well-defined morphological problem while testing the underlying lexicographic model.

1 https://www.perseus.tufts.edu/hopper/morph
2 https://logeion.uchicago.edu/morpho/%CE%BB%CF%8C%CE%B3%CE%BF%CF%82
3 https://logeion.uchicago.edu/
4 https://www.perseus.tufts.edu/hopper/
5 https://www.perseus.tufts.edu/hopper/text?doc=Perseus:text:1999.04.0057

From the user's perspective, the prototype processes each query through a sequential workflow. When an inflected word form is entered, the system first attempts to match it against the curated dataset. If a match is found, the corresponding lemma and morphological analysis are returned directly. If no match is found, the input is passed to an automatic morphological analyser, which generates one or more candidate lemmas that can be linked to dictionary entries.

When the user enters a form, the prototype provides three possible types of results. A GOLD result is returned when the input form is attested in the curated dataset. An ACCEPTABLE result is returned when the form is not present in the dataset but the automatic analysis proposes a lemma that is included in the controlled inventory. An EXTERNAL result is returned when the automatic analysis proposes a lemma that is not attested in the dataset. These latter results are explicitly marked as automatically generated. This classification makes the origin of the result explicit, allowing users to distinguish between data derived from the curated dataset and data generated through automatic analysis, and thus to navigate dictionary access more transparently.

From a technical point of view, the prototype consists of two main components: a curated dataset of inflected forms linked to their lemmas and their morphological analysis, and an automatic morphological analyser, specifically Morpheus, used as a fallback. The dataset is stored as a structured table, while the analyser is queried only when no match is found in the dataset. The prototype has been implemented in Python, with Unicode normalization for polytonic Greek and a simple web-based interface developed in HTML.

This prototype demonstrates how a form-to-lemma access model can be applied to Ancient Greek, combining curated lexical data and automatic analysis to enable inflected-form search while preserving a clear distinction between different types of results. It is currently implemented as a local system and is not yet integrated into an online dictionary environment. Future work will focus on integrating the model into a fully functional digital dictionary and testing its application on a larger lexical inventory.

References

Crane, G. (1991). Generating and Parsing Classical Greek. *Literary and Linguistic Computing*, 6(4), 243–245.

Crane, G. (1996). Building a Digital Library: The Perseus Project as a Case Study in the Humanities. *D-Lib Magazine*.

Logeion. Retrieved May 2026, from https://logeion.uchicago.edu/

Perseus Digital Library. Retrieved May 2026, from https://www.perseus.tufts.edu/hopper/

Perseus Digital Library: Morphological Analysis Tool. Retrieved May 2026, from https://www.perseus.tufts.edu/hopper/morph

Contact information

Elena Castaldi
Erasmus Mundus Joint Master in Lexicography
castaldi.elena2003@gmail.com

Elina Chadjipapa, Zoe Gavriilidou, Maria Mitsiaki & Ifigeneia Dosi

CHATBOTS IN PEDAGOGICAL LEXICOGRAPHY

Evaluating AI-Generated Dictionary Entries from Three LLMs

Keywords Helix; pedagogical lexicography; AI; ChatGPT; Gemini; DeepSeek; dictionary entries

As early as the eighteenth century, dictionary compilation has been perceived as a labor intensive and time-consuming endeavor, often characterized as meticulous and repetitive work (Lew, 2023). The creation of lexicographic articles, in particular, constitutes a complex process requiring the careful coordination of multiple parameters, including the target user group, users' native language, the type and purpose of the dictionary, as well as the design of its macrostructure and the organization of its microstructure (Hartmann, 2001). In pedagogical lexicography, this complexity is further intensified, as definitions and supporting information must also account for learners' age, language proficiency, learning objectives, and the conditions under which the dictionary is consulted, such as reception- or production-oriented use (Tarp, 2011). Against this background, contemporary lexicography has increasingly emphasized the need to reduce the time and cost of dictionary making, with automation widely regarded as a key factor in achieving this goal (Jakubíček & Rundell, 2023; Rundell, 2024). Recent advances in artificial intelligence, particularly over the past five years, have therefore prompted lexicographers to explore new solutions for both dictionary compilation and use. Notably, the emergence of Large Language Models (LLMs) has intensified interest in the potential role of chatbots in supporting—and possibly reshaping—various stages of the lexicographic process, including the generation and mediation of lexicographic articles (Phoodai & Rikk, 2023; Rees & Lew, 2023; Lew et al., 2024; Li & Tarp, 2024; Ptasznik et al., 2024; Ptasznik & Lew, 2025).

Against this backdrop, the present study investigates how three widely used LLMs generate dictionary entries and examines their potential contributions and limitations within pedagogical lexicography. Focusing on AI-generated dictionary entries produced by ChatGPT, Gemini, and DeepSeek, the study adopts a comparative, criteria-based evaluation framework grounded in learner lexicography, pedagogical principles, and AI-assisted language learning. Its primary objective is to assess the extent to which LLMs can simulate the writing practices and decision-making processes of human lexicographers. To this end, the models' ability to generate definitions, examples, and idiomatic expressions is evaluated using twenty carefully selected entries from Helix, a bilingual illustrated lexicon designed for Greek heritage learners (Chadjipapa & Gavriilidou, 2024). All models were prompted under controlled conditions to produce dictionary entries including definitions, sense distinctions, usage notes, examples, collocations, register labels, and, where relevant, learner-oriented or cross-linguistic explanations.

The chatbot-generated lexicographic articles were evaluated by four expert lexicographers specializing in pedagogical lexicography. The evaluation combined qualitative and quantitative dimensions and was conducted under blind conditions, with assessors unaware of whether entries were produced by human lexicographers or AI systems. Entries were assessed along five main axes: (a) definitional accuracy and semantic precision, (b) internal consistency and sense discrimination, (c) pedagogical usefulness for learners, (d) transparency and explicability of lexical information, and (e) alignment with established learner lexicographic principles, including clarity, economy, controlled defining vocabulary, and relevance to communicative use. Particular attention was paid to whether the models reproduced established lexicographic

conventions, extended them, or violated core norms. The results show that all three LLMs are capable of producing superficially well-formed dictionary entries that resemble the structure and tone of learner dictionaries. Nevertheless, systematic differences were observed. ChatGPT demonstrated stronger coherence in extended explanations and pedagogical scaffolding, particularly in usage notes and learner oriented paraphrasing. Gemini excelled in the range of examples and contextual variation, though sometimes at the expense of semantic precision, while DeepSeek offered comparatively limited learner-specific guidance. These findings align with patterns reported in previous research (Lew, 2024). Across models, recurrent weaknesses included inconsistent sense hierarchies, blurred distinctions between core meaning and pragmatic inference, occasional hallucinated usage restrictions, and limited sensitivity to frequency, register, and learner difficulty.

Despite these limitations, the findings point to a significant pedagogical opportunity: when guided by appropriate pedagogical prompts and lexicographic constraints, LLMs can meaningfully support aspects of dictionary compilation. The contribution of this paper is therefore threefold. First, it provides one of the first systematic and comparative evaluations of AI-generated lexicographic articles produced by multiple LLMs, employing a blind assessment procedure that enables a more nuanced evaluation of whether chatbot generated outputs meet established lexicographic standards beyond surface-level fluency. Second, it proposes an evaluative framework grounded in pedagogical lexicography, offering concrete criteria for assessing and improving chatbot-mediated dictionary entries, particularly with respect to learner-oriented definitions, sense structuring, and usage information. Third, the paper advances the methodological integration of chatbots into lexicographic workflows by examining prompt design as a key mediating factor in AI-assisted dictionary making. Focusing on few-shot prompting (Brown et al., 2020), it analyes how chatbots adapt their outputs in response to carefully curated lexicographic examples and identifies parameters necessary for the effective, reliable, and pedagogically sound use of prompts. Taken together, these contributions frame chatbots as potential partners in pedagogical lexicography, capable of extending traditional dictionary functions through interaction, personalization, and AI supported mediation when guided by lexicographic expertise.

References

Brown, TB., Mann. B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G. Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan & H. Lin (Eds.), *Proceedings of the 34th International Conference on Neural Information Processing Systems* (NIPS '20, pp. 1877–1901). Curran Associates, Inc.

Chadjipapa, E. & Gavriilidou, Z. (2024). Helix – A Bilingual Illustrated Dictionary for Greek Heritage Learners, In K. Š. Despot, A. Ostroški Anić & I. Brač (Eds.) *Lexicography and Semantics, Proceedings of the XXI EURALEX International Congress* (pp. 566-575). Institut za hrvatski jezik.

Jakubíček, 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. Jakubíček (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. (2024). Dictionaries and lexicography in the AI era. *Humanities and Social Sciences Communications*, 11, 426. https://doi.org/10.1057/s41599-024-02889-7

Lew, R., Ptasznik, B. & Wolfer, S. (2024). The effectiveness of ChatGPT as a lexical tool for English, compared with a bilingual dictionary and a monolingual learner's dictionary. *Humanities and Social Sciences Communications,* 11, 1324. https://doi.org/10.1057/s41599-024-03775-y

Li, Q., & Tarp, S. (2024). Using Generative AI to Provide High-Quality Lexicographic Assistance to Chinese Learners of English. *Lexikos*, 34(1), 397-418. https://doi.org/10.5788/34-1 1944

Phoodai, C. & Rikk, R. (2023). Exploring the Capabilities of ChatGPT for LexicographicalPurposes: A Comparison with Oxford Advanced Learner's Dictionary within the Microstructural Framework, In M. Medveď, M. Měchura, C. Tiberius, I. Kosem, J. Kallas & M. Jakubíček (Eds.), *Proceedings of the eLex 2023 Conference: Electronic Lexicography in the21st Century* (pp. 335–365). Lexical Computing.

Ptasznik, B., & Lew, R. (2025). Dictionaries versus AI Tools through the Eyes of English Majors. *International Journal of Lexicography*, 38(2), 140 158. https://doi.org/10.1093/ijl/ecaf005

Ptasznik, B., Wolfer, S., & Lew, R. (2024). A Learners' Dictionary Versus ChatGPT in Receptive and Productive Lexical Tasks. *International Journal of Lexicography*, 37(3), pp. 322–336.

Rees, G.P., & Lew, R. (2023). 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.

Rundell, M. (2024). Automating the creation of dictionaries: Are we nearly there? *Humanising Language Teaching*, 26(1).

Tarp, S. (2011). Pedagogical Lexicography: Towards a New and Strict Typology Corresponding to the Present State-of-the-Art. *Lexikos*, 21, 217-231.

Contact information

Elina Chadjipapa
Democritus University of Thrace
elinaxp@hotmail.com

Zoe Gavriilidou
Democritus University of Thrace
zoegab@otenet.gr

Maria Mitsiaki
Democritus University of Thrace
mitsiakim@gmail.com

Ifigeneia Dosi
Democritus University of Thrace
dosi.if@gmail.com

John J. Chew III, Emma Ferrett & Anastasia Riehl

CANADIAN ENGLISH DICTIONARY

A Progress Report

Keywords Lexicography; Canadian English; online lexicography

Canada is long overdue for a new English dictionary. In this presentation, we describe progress to date on the *Canadian English Dictionary* (CED), a new general dictionary of Canadian English being edited by the recently incorporated Society for Canadian English, with the support of Editors Canada, the University of British Columbia's Canadian Word Centre, and Queen's University's Strathy Language Unit. A brief examination of the last generation of Canadian English dictionaries (*Gage Canadian Dictionary* [1996], *ITP Nelson Dictionary* [1997], *Canadian Oxford Dictionary,* 2nd edition [2004]) reveals the need for an up-to-date resource. Oxford's second (and last) edition, published in 2004, remains the most recent resource for Canadian dictionary users, indicating that over twenty years of language change remains undocumented.

Currently, Canadian English speakers must turn either to American or British English dictionaries for up-to-date lexicographical information. We need look no further than a recent spelling controversy that flared up in the media to remind ourselves that Canada's variety of English is distinct, and that it matters to the population: our prime minister's use of British *-ise* spellings in government documents has been met with criticism by Editors Canada, who remind readers that Canadians use *-ize* spellings (see Yousif 2025; 'Linguistic experts' 2025) and that Canadian English spelling conforms neither to American English nor British English (e.g. *Canadian tire centre* vs. *British tyre centre* vs. *American tire center*). Such cases reveal the pressing need for a current resource that lays out and cements our own variety of spellings, pronunciations, usages, and unique lexical items, as distinct from the British and American varieties. Additionally, during this time of geopolitical uncertainty, where Canada's identity and even sovereignty feel under threat, Canadians want and need a resource to call their own.

In 2022, Editors Canada recruited John Chew as editor-in-chief of the project. The core team quickly expanded, bringing in linguists, editors, lexicographers, fundraisers, and social media professionals, then set about documenting what would be involved in creating our new dictionary, its style guides, its procedures, its business model, its governance, and its public engagement. We chose as a starting point and proof-of-concept, fascicle Q. In Canadian English, Q is a significant letter because it demonstrates well our need for an up-to-date, inclusive dictionary: the definientia for *queen bee*, *quantum computing*, and *quarantine*, for example, have gone through major transformations since 2004 and need addressing; neologisms (e.g., *quiet quitting*, *quadrex*) need documenting; now offensive terms need labelling (e.g., *quadroon*, *quarter-blood Indian*); Canadian legal and military terms that carry some history from either American and British systems, and yet correspond fully with neither (e.g., *Question Period*, *question of law*, and *quartermaster*), need distinguishing; and the important work of decolonizing Indigenous language spellings and definientia (e.g., *Q'omoχws*, *Queen Charlotte Island caribou*) is urgent as Canada pursues reconciliation with Indigenous Peoples.

With these goals in mind, we discuss our establishment of lexicographical principles using several entries of interest from within the Q fascicle. We reveal how our editors construct all the aspects of an entry, with attention paid to issues of definition, spelling, pronunciation, etymology, labels, and presentation. We provide an overview of methodology, including the compilation of the word list, editorial procedures, drafting of editorial guides, and corpus construction.

We also discuss our social media and outreach efforts, including the recent establishment of a Canadian Word of the Year, with which we hope to bring attention to Canadian English lexica amongst other major British and American dictionaries and their respective Word of the Year traditions. The very first Canadian Word of the Year, *maplewash* (the deceptive practice of making things seem more Canadian than they actually are) gained popularity in 2025 as Canadians embraced local products in the face of US tariffs. Its selection as Word of Year in a public poll indicates Canadians' desires to establish our own ideological and linguistic identity on the global stage.

The initial fascicle Q, consisting of 500+ entries, is now available for online review, and work is continuing alphabetically, currently in fascicle R. If our ongoing funding efforts are successful, we plan to offer a free app by 2028, with premium and institutional subscriptions to ensure the sustainability of the CED. We look forward to introducing our project to international lexicographical experts and we welcome informed feedback on our methodologies and draft entries.

References

Avis, W. S. (1967). *Gage Dictionary of Canadian English: The Senior Dictionary*. Gage.

Barber, K. (Ed.) (2004). *The Canadian Oxford Dictionary*. Oxford University Press Canada.

Chew, J. J. III (Ed.) (work in progress). *The Canadian English Dictionary* (development edition). Retrieved from www.canadianenglishdictionary.ca/dev

Harkness, J., Friend, D., Keefer, J., Liebman, D. & Sutherland, F. (Eds.) (1997). *ITP Nelson Canadian Dictionary of the English Language. An Encyclopedic Reference*. ITP Nelson.

Linguistic experts urge Carney government to stop using British spellings. (2025). CBC.ca. Retrieved Jan 27, 2026, from https://www.cbc.ca/news/canada/carney-criticized-for-british-spelling-9.7015702

Pratt, T.K. (Ed.) (2013). *Nelson Gage Canadian Paperback Dictionary*. Nelson Education Ltd.

Yousif, Nadine. (2025). *Canada's Carney called out for 'utilizing' British spelling*. BBC. Retrieved Jan 30, 2026, from https://www.bbc.com/news/articles/cj69d89l8l5o

Contact information

John J. Chew, III
Canadian English Dictionary
info@canadianenglishdictionary.ca

Emma Ferrett
Queen's University
emma.ferrett@queensu.ca

Anastasia Riehl
Queen's University
riehla@queensu.ca

Luisa Cimander & Petra Storjohann

TRACING NEOLOGISMS IN NEO-DISCOURSES

Diachronic and Cross-Domain Perspectives

Keywords Neologisms; neo-discourses; COVID-19 discourse; user access in online dictionaries; semantic categorisation; Neo2020+; AI support for lexicographic work

Documenting the surge of new lexical items during the COVID-19 pandemic presented a challenge worldwide (Klosa-Kückelhaus & Kernermann, 2024). German Neologisms emerging between 2020 and 2023 were collected (Klosa-Kückelhaus, 2021) and incorporated into a dedicated section of the *Neologismenwörterbuch* (2006ff.), initially as a simple word list. Each entry includes a brief paraphrase and illustrative usage examples, often drawn from online sources, as validated corpus data were not yet available at the time of compilation. This section now comprises about 2,500 lemmas (*Neuer Wortschatz rund um die Coronapandemie*, 2020ff.). Taken together, these document the rise of a neo-discourse marked by an influx of new words, structural complexity, lexical competition as well as shifting narrative and topical emphases (Storjohann & Cimander, 2022, Mattfeld et al., 2025).

Consequently, the need arose to enhance user access to these entries. This initially involved assigning ontological categories through a triple-blind procedure, thereby establishing domains and semantic groupings for the lexical items. These were taken from the existing domain classes of the *Neologismenwörterbuch* (2006ff.) used for grouping headwords between 1990 and 2020. The aim was to create the kind of semantic and thematic connectivity that is crucial for online dictionaries (Dornseiff, 2020). In addition to general domains such as ECONOMY, EDUCATION, HEALTH/MEDICINE, POLITICS (cf. Ilić, 2021), a domain specific to Covid discourse was identified: MEASURES encompasses terms that refer to regulatory interventions and restrictions during the pandemic.

A representation of Covid-related vocabulary is currently being integrated into the new platform IDS *Neo2020+*, as it offers a range of entry types, e.g. single entries, discourse entries and discourse-morphological entries (Storjohann, in prep.). These enable the documentation of neologisms in single headword entries, in clusters, by concept, by domain and across temporal stages. This structure allows users to explore individual lexical items, their diachronic development, functions in the respective neo-discourse and semantic relations, especially among synonyms. Consequently, users may consult either individual COVID-related lexical items or, alternatively, a discourse entry that aggregates COVID-19 neologisms according to various criteria and presents them in relation to temporal, thematic, or structural dimensions.

While implementing new formats, a series of experiments will assess the extent to which AI can support or complement lexicographic work. These include AI-based categorisation of terms compared to manual classification, the identification of synonyms and the determination of first attestations (Storjohann & Cimander, 2022). AI may also facilitate the migration of the concise entries into the revised and more complex architecture of new headword entries. This step is necessary to construct a comprehensive discourse entry, "COVID," by incorporating and elaborating upon details that are integral to individual COVID-related terms. Given the large number of items, enrichment through AI-assisted methods appears promising, e.g. by adding reliable citations, encyclopaedic or historical context, information on word-formation productivity, collocation profiles. It will also be examined whether such information can not only be

identified by AI but also automatically assigned to designated elements at specific positions within larger XML-based entries. Positive outcomes could substantially influence the scope and depth of the documentation of Covid neologisms and, more broadly, the modelling of this neo-discourse within the new dictionary.

References

Dornseiff, F. (2020[9]). *Der deutsche Wortschatz nach Sachgruppen*. De Gruyter.

Ilić, K. (2021). *Klassifizierung des neuen Wortschatzes rund um die Coronapandemie nach Sachgruppen* [Master's thesis, University of Belgrad].

Klosa-Kückelhaus, A. (Ed.) (2021). *Sprache in der Coronakrise. Dynamischer Wandel in Lexikon und Kommunikation*. IDS-Verlag.

Klosa-Kückelhaus, A. & Kernerman, I. (Eds.) (2024). *Lexicography of Coronavirus-related Neologisms.* De Gruyter.

Mattfeldt, A., Warnke, I. H. & Herford, L. (2025). Neodiskurse und ihre Morphologie – Bemerkungen zum Substantivprimat. In Michel, S. (Ed.), *Diskursmorphologie: Ansätze und Fallstudien zur Schnittstelle zwischen Morphologie und Diskurslinguistik* (pp. 121–152). De Gruyter. https://doi.org/10.1515/9783111363936-005

Neologismenwörterbuch (2006ff.). Retrieved May 27, 2025, from https://www.owid.de/docs/neo/start.jsp. Leibniz-Institut für Deutsche Sprache.

Neuer Wortschatz rund um die Coronapandemie (2020ff.). Retrieved May 27, 2025, from www.owid.de/docs/neo/listen/corona.jsp. Leibniz-Institut für Deutsche Sprache.

Storjohann, P. (in prep.). The interplay between individual new lexical items and multi-layered discourse representations of neologisms. In Ph. Stöckle, W. Koppensteiner, K. Korecky-Kröll, M. L. Krapp & T. Wissik (Eds.), *Lexicography in the Age of AI. Proceedings of the XXII EURALEX*. Austrian Academy of Sciences Press.

Storjohann, P. & Cimander, L. (2022). Annäherung an den Coronadiskurs in der öffentlichen Kommunikation mithilfe von Neologismen. In M. Jakosz & M. Kałasznik (Eds.), *Corona-Virus-Pandemie: Diverse Zugänge zu einem aktuellen Superdiskurs* (pp. 25–50). V&R Unipress.

Contact information

Luisa Cimander
Leibniz Institute for the German Language, Mannheim
cimander@ids-mannheim.de

Petra Storjohann
Leibniz Institute for the German Language, Mannheim
storjohann@ids-mannheim.de

Václav Cvrček, Martin Stluka & Klára Pivoňková

TRACKING EIGHT CENTURIES

The Czech Monitor Corpus as a Resource for Diachronic Lexical Research

Keywords Monitor Corpus; Timeline application; diachronic lexical research; Czech language

This paper demonstrates the central role of richly annotated diachronic corpora in investigating historical lexical change, using the newly compiled Monitor Corpus of Czech as a case study. The corpus covers over eight centuries of the Czech written tradition and is designed as a general-purpose, consistently annotated resource for diachronic research. Unlike earlier diachronic corpora of Czech, the Monitor Corpus aims at systematic genre balance (where historical conditions permit) and at comparability of annotation across periods. To make the corpus maximally exploitable, it is accompanied by a dedicated application, Timeline, which supports queries and visualizations tailored to diachronic analysis.

The design of the Monitor Corpus seeks to represent, for each period, the broadest possible range of text types at the most general level: fiction, non-fiction, and journalism. Of course, this ideal cannot be met uniformly throughout the history of Czech. For some periods, no specialist texts have survived in Czech and journalism emerges only from the 18th century onwards. Nevertheless, in those historical stages where it is feasible, the corpus includes roughly comparable proportions of fiction, non-fiction, and journalistic writing, thereby enabling meaningful comparisons of lexical developments across time and genre.

For lexicographical and broader linguistic research, uniform, transparent and, above all, consistent annotation over time is essential. A key methodological challenge is how to handle both graphical and morphological variability so that it becomes possible to track lexical items across centuries. For example, the adjective *estetický* 'aesthetic' is attested in the 19th-century material in numerous spelling variants (e.g. *aestetického*, *Esthetickému*, *aesthetickou*). These forms must be recognized as instances of a single lemma in order to study its history in a principled way. To address this, we compiled three manually annotated etalon corpora representing different historical stages of Czech and used them as training data for a lemmatizer and POS/morphological tagger developed within the Universal Dependencies (UD) framework (de Marneffe et al., 2021; Zeman et al., 2021). The result is a diachronic corpus that is lemmatized and morphologically annotated according to a single, coherent scheme.

The benefits of this annotation become especially clear when investigating developmental tendencies that would be difficult or impossible to access using only raw text. By systematically linking historically and graphically divergent word forms under a single lemma and assigning each token a detailed morphological description, we can reliably estimate frequencies, identify collocational patterns, and detect changes in distribution across grammatical categories and constructions. The paper will describe how historical variants were handled in the annotation workflow, and how the resulting data can be exploited to study lexical change.

As concrete case studies illustrating how the Monitor Corpus and Timeline application can be used in practice, we focus on examples of nouns that have undergone semantic change or diachronic shifts in usage, such as *nápad* 'attack'/'idea', *puška* 'container'/'rifle', and *národ* 'family'/'nation'. For example, the word *národ* has undergone a complex semantic evolution since the earliest periods of Czech: attested meanings include 'fruit', 'family/kin', 'pagans', and progressively the now dominant sense 'community of people sharing a territory or language'.

Its sociopolitical salience in the 19th century is reflected in a marked increase in frequency in that period. The annotated corpus allows us to examine not only this overall frequency development, but also changes in the distribution of grammatical forms and constructions, and to relate them to semantic and discursive shifts.

In Old Czech, plural forms are strongly predominant, and one important meaning of the time—'pagans'—is expressed exclusively in the plural (*národové* 'nations'). From the early 19th century onward we observe a significant rise in genitive forms, and the lexical content of národ becomes increasingly abstract and context-dependent. The heads of noun phrases in which *národ* appears make an important contribution to its discursive interpretation. In the 19th century, typical genitive constructions include *duch národa* 'the spirit of the nation', and *rozkvět národa* 'the flourishing of the nation', reflecting nationalist and Romantic discourses. By contrast, in 21st-century texts, constructions such as *společenství národů* 'community/commonwealth of nations' and *organizace národů* 'organization of nations' become prominent, mirroring a shift towards international and institutional frames.

Using a small set of nouns as case studies, the paper will show how lemmatization enables the automatic extraction of statistically robust lemma-based collocations across periods and genres, and how morphological annotation supports the identification of typical semantic–syntactic patterns. Both lines of research use the Timeline application, which allows various frequency data and collocation profiles to be extracted and visualized on a timeline. The Monitor corpus and the Timeline application are currently in beta and are expected to be made publicly available by the end of 2026.

References

De Marneffe, M.-C., Manning, C. D., Nivre, J., & Zeman, D. (2021). Universal dependencies. *Computational Linguistics*, *47*(2), 255–308. https://doi.org/10.1162/coli_a_00402

Zeman, D., Kosek, P., Březina, M., & Pergler, J. (2023). Morphosyntactic annotation in universal dependencies for old czech. *Jazykovedný časopis/Journal of Linguistics*, *74*(1), 214–222.

Contact information

Václav Cvrček
Charles University – Faculty of Arts
vaclav.cvrcek@ff.cuni.cz

Martin Stluka
Charles University – Faculty of Arts
martin.stluka@ff.cuni.cz

Klára Pivoňková
Charles University – Faculty of Arts
klara.pivonkova@ff.cuni.cz

Acknowledgements

The project HiČKoK (History of Czech in Corpus Continuum, see https://korpus.cz/hickok) is supported by the Technology Agency of the Czech Republic within the Programme for Support of Applied Research and Innovation SIGMA (no. TQ01000072).

Veronique De Tier, Katrien Depuydt, Koen Mertens & Mathieu Fannee

FROM PILOT TO PRACTICE

Two Years of Progress in Developing a Shared Infrastructure for Dutch Dialect Dictionaries

Keywords Dutch dialect lexicography; digital lexicography; lexical infrastructure; dialect documentation; lexical databases; participative lexicography

At Euralex 2024, the initial design of a digital lexical infrastructure for Dutch Dialects was presented, developed in response to requests from dialect organizations in the Netherlands and Flanders to support the compilation and publication of dialect dictionaries. That contribution focused primarily on conceptual design choices and technical feasibility. Two years later, the project has moved from exploration to practice. The infrastructure has been implemented, tested, and refined through real-world use by two dialect communities. This paper reports on these developments, with particular attention to lexicographic equivalence, source integration, orthographic variation, and the role of non-professional lexicographers in collaborative dialect dictionary making.

Despite ongoing dialect loss, public engagement with regional varieties of Dutch remains strong. Dialect associations play a central role in documentation, education, and dissemination, often through dictionary projects or learning materials. This work is largely carried out by volunteers. While their commitment and local knowledge are indispensable, such initiatives typically operate with limited resources and without structural access to linguistic or lexicographic training. As a result, dialect dictionaries are often developed in isolation, follow divergent conventions, and are difficult to update, compare, or integrate. Lexical data are also at risk of being lost due to inadequate long-term preservation (Van Keymeulen et al. 2009). The project discussed here addresses these challenges by developing a shared digital infrastructure that supports sustainable dialect lexicography while accommodating differences in expertise, goals, working practices, and community-specific preferences.

Two pilot varieties were selected: Bildts, a Frisian-Dutch contact variety spoken in the province of Fryslân (Friesland), and West-Overijssels, a regional variety with a long lexicographic tradition. These cases were chosen not only for their linguistic characteristics, but also for the presence of active, volunteer-driven language communities. In both pilots, the first step consisted of establishing a lexical database as the backbone for future dictionary products, reflecting the lexicographic principle that a well-defined macrostructure is a prerequisite for consistent and reusable microstructural description.

The editorial workflow is centred on a lexical data editing environment (Lex'it) (fig. 1) that supports structured data entry, linking, and revision. As an initial alignment device, a Dutch base word list derived from frequency-oriented learner resources was used. This list functions as a pragmatic onomasiological access structure, intended to facilitate dialect acquisition, translation activities, and comparison across dialects. Experience from the pilots confirmed, however, that frequency-based lists alone are insufficient for representing dialect lexicons. Many high-frequency standard-language words lack clear dialect equivalents, while culturally salient dialect words are often absent from such lists. Volunteers frequently prioritize local relevance and expressive richness over frequency.

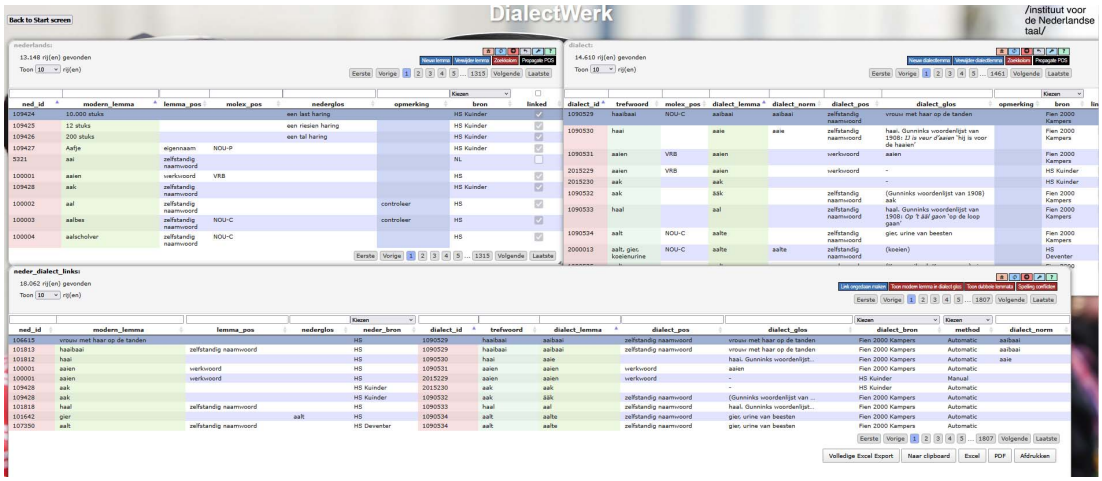


FIGURE 1 Lexical Data Editing Environment (West-Overijssels)

In both pilots, existing dialect dictionaries (Buwalda 2013, Fien 2000, Kamman 1990, Kuijk & Van Baalen 2015) were integrated into the editing environment. For Bildts, one authoritative dictionary served as the primary source, supplemented with newly attested items stored in an extensible lexicon. For West-Overijssels, data from three independent dictionaries were combined, enabling the compilation of a compact, learner-oriented pocket dictionary suitable for a broader region. This process foregrounded classical lexicographic issues such as source criticism, sense delimitation, synonymy, and the treatment of variants.

Combining multiple sources substantially increased lexical coverage but also introduced ambiguity. Differences in spelling conventions, semantic scope, and levels of detail between dictionaries required explicit editorial decisions. Orthographic variation proved particularly prominent in volunteer-based projects, as contributors often follow different local, historical, or personal spelling systems. The infrastructure therefore supports the coexistence of multiple spelling variants without enforcing premature standardization, while still providing reliable search and access mechanisms (fig. 2).

The pilots further demonstrated that lexical linking cannot be reduced to simple one-to-one equivalence. Partial equivalence, context-dependent meanings, and culture-specific concepts occur frequently in dialects. All editorial actions are logged within the system, enabling transparency, revision, and quality control (fig. 3).

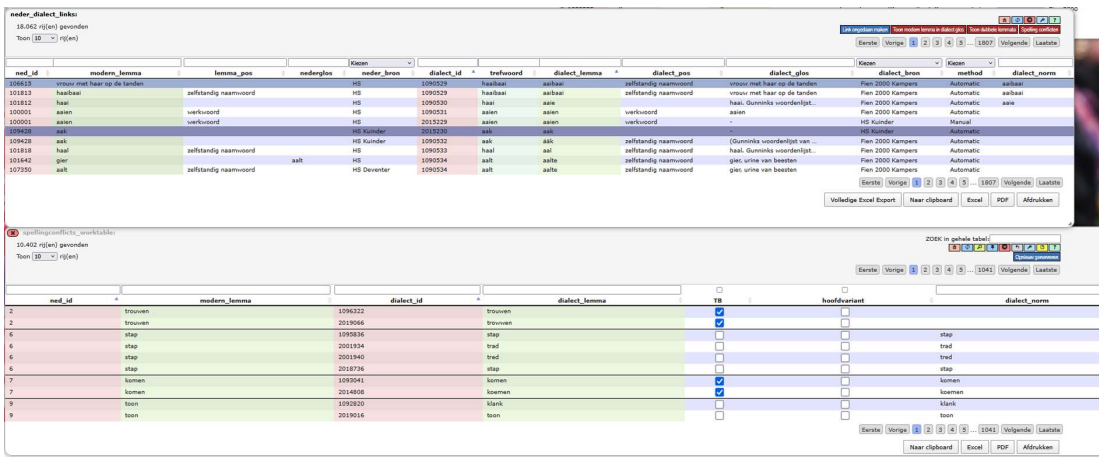


FIGURE 2 Spelling normalisation

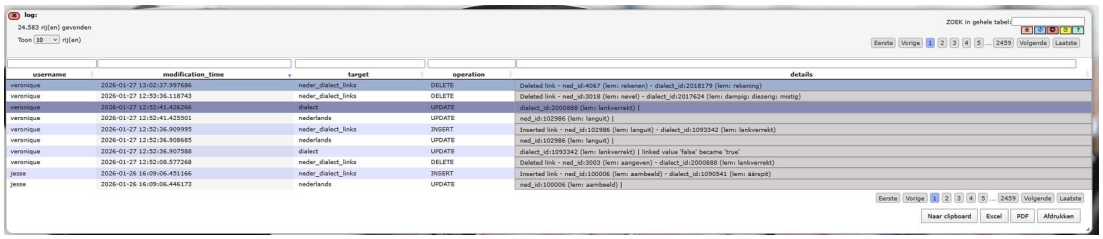


FIGURE 3 Log system

FIGURE 4 Search facilities of the Bildts pilot

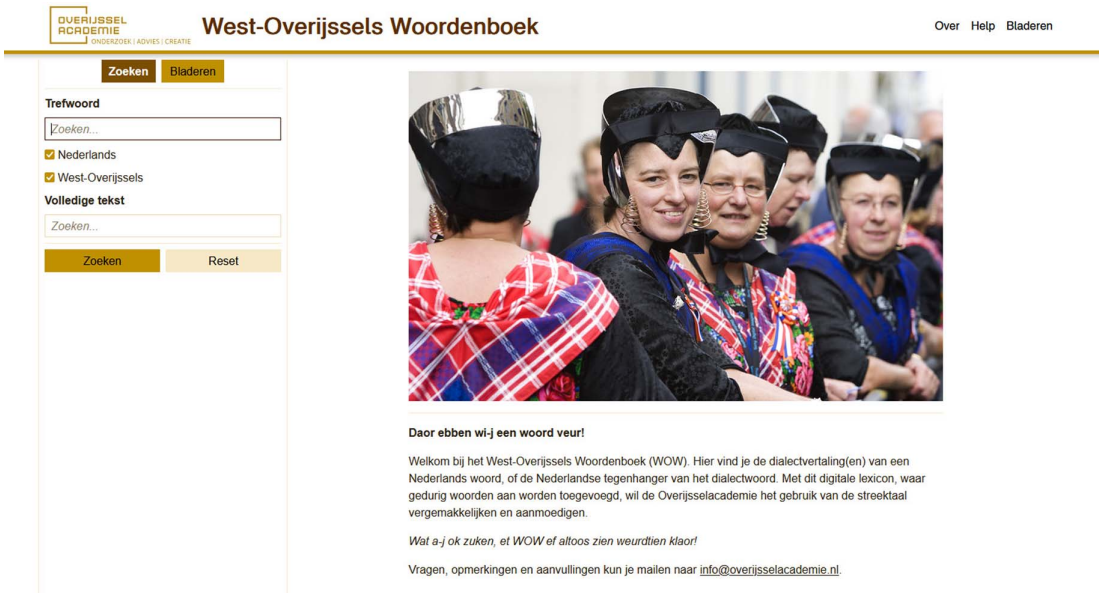


FIGURE 5 The website and search facilities of the West-Overijssels Woordenboek (https://wow.ivdnt.org/sites/overijssels/)

A key outcome after two years is that the same underlying system now supports two distinct public dictionary products (e.g. fig. 4 and 5). The Bildts and West-Overijssels lexica are published on the respective websites of the collaborating organizations, each with its own interface and local identity, reflecting heterogeneous user groups ranging from experienced dialect

speakers to beginners. At the infrastructural level, however, both products are generated from the same database and editorial workflow, illustrating a clear separation between lexicographic data, structure, and presentation.

The pilots show that creating digital dialect dictionaries involves not only linguistic data, but also volunteer practices, orthographic diversity, and diverse user needs. By embedding lexicographic theory into infrastructural design, the project offers a sustainable model that combines scholarly rigor with accessibility.

References

Buwalda, H. S., Buwalda, S. H., Van der Burg, A. C. B., & Sijens, H. (2013). *Woordeboek fan 't Bildts: Tweede ferbeterde en anfulde edisy* (2e ed.). Fryske Akademy.

Fien, A., Bloemhoff-de Bruijn, P. C. G. M., & Gunnink, J. (2000). *Woordenboek van de Kamper taal*. IJsselacademie.

Hazenberg, S., & Hulstijn, J. H. (1992). Woorden op zicht: Woordselectie ten behoeve van het NT2-onderwijs. *Levende Talen Magazine*, 79(467), 2–7.

Kamman, R. (1990). *Woordenboek van het Kuinders*. IJsselacademie.

Kuijk, G., & Van Baalen, H. (2015). *Idioom Deventers*. Eigen beheer.

Van Keymeulen, J., & Oosterhof, A. (2009). Local dialect dictionaries and a proposal for a dictionary of the Dutch dialects. In A. Lenz, C. Gooskens, & S. Reker (Eds.), *Low Saxonian dialects across borders: Synchrony and diachrony* (pp. 109–124). Franz Steiner.

West-Overijssels Woordenboek. Retrieved from https://wow.ivdnt.org/sites/overijssels/

Contact information

Veronique De Tier
Dutch Language Institute
veronique.detier@ivdnt.org

Katrien Depuydt
Dutch Language Institute
katrien.depuydt@ivdnt.org

Koen Mertens
Dutch Language Institute
koen.mertens@ivdnt.org

Mathieu Fannee
Dutch Language Institute
mathieu.fannee@ivdnt.org

Kristina Š. Despot, Ana Ostroški Anić & Matej Klemen

BEYOND BINARY METAPHOR DETECTION

Dataset-Driven Evaluation of Conceptual Metaphor Type Recognition in Croatian and Slovene

Keywords Metaphor dataset; metaphor detection; natural language inference; semantic similarity scoring; LLMs

Large language models (LLMs) continue to perform suboptimally in tasks involving metaphor detection and explanation (Devlin et al., 2019; Liu et al., 2022; Klemen & Robnik Šikonja, 2023, Kim et al., 2023; He et al., 2023; Despot, Ostroški Anić, & Veale, 2023; Tong et al., 2024; Puraivan et al., 2024; Yang et al., 2025; Chen & Wang, 2025). Core challenges lie in the inherent semantic complexity of metaphorical expressions, and in the dearth of structured metaphor corpora. In response to this limitation, we introduce CroSloMet, a novel structured metaphor dataset for two South Slavic languages, Croatian and Slovene, designed to support both metaphor identification and explanation tasks (Ge et al., 2023, Despot et al., 2025). CroSloMet comprises over 1,120 metaphorical sentences and 1,120 matched literal sentences for each language, annotated with semantic metadata. Each data point is aligned with corresponding conceptual metaphors, multi-word expressions, canonical forms, and literal usages, providing a multi-layered annotation scheme that goes beyond surface metaphor tagging to encode conceptual and linguistic structure. This resource is grounded in the MetaNet.HR framework (Despot et al., 2019) enabling researchers to explore generality and specificity in metaphoric expressions. The dataset supports metaphor identification tasks—the classification of sentences as metaphorical vs. literal through annotated pairs that preserve lexical overlap between figurative and non-figurative contexts. Moreover, it facilitates metaphor explanation (conceptual metaphor type recognition, where the conceptual metaphor name is used as a proxy for explanation), allowing evaluation of model output not merely on binary classification but on the degree to which generated conceptual metaphor type capture intended conceptual mappings. CroSloMet's dual annotation quality makes it suitable for corpus-based analysis, cognitive linguistics studies, and the development of metaphor understanding modules in NLP systems. To demonstrate the dataset's utility, the original work reports preliminary experiments using a fine-tuned CroSloEngual BERT model for metaphor classification, achieving an accuracy of 88.5%, and an evaluation of LLaMA 3-8B for metaphor detection. While classification results were promising, strict exact-match evaluation for explanation generation (conceptual metaphor type recognition) yielded low scores, revealing a gap between model outputs and human interpretive expectations despite the semantic validity of many generated texts. This discrepancy pointed to the need for improved evaluation metrics that capture semantic similarity and interpretive nuance.

Building on these insights, the current paper proposes to extend CroSloMet with a multi-level validation framework based on the comparison of manual annotation to natural language inference (NLI), semantic similarity scoring, and LLM-based judgment to assess metaphor explanations more holistically. By integrating human expert judgments with automatic evaluation signals, the proposed research seeks to close the gap between computational performance and cognitive plausibility in metaphor interpretation. This study leverages this dataset and evaluation framework to address several core research questions: How can structured metaphor annotations be exploited to improve cross-lingual metaphor understanding? What evaluation strategies best capture the quality of metaphor explanations in LLM-generated text? And how can conceptual metaphors be operationalized to support both symbolic and

distributional approaches to figurative language modelling? To answer these questions, the paper outlines an experimental pipeline that integrates CroSloMet with advanced representation learning techniques, such as transformer architectures and semantic embedding spaces calibrated for figurative meaning. A key aspect of this research is the incorporation of semantic generalization layers within explanation models that can abstract over lexical variation to capture deep conceptual mappings. We take advantage of the hierarchical organization of conceptual metaphors in the MetaNet.HR database (Despot et al., 2019) to refine the evaluation of model-generated metaphor explanations. Instead of relying solely on exact matches between predicted and gold-standard conceptual metaphors, we implement a graded evaluation scheme that accounts for metaphor hierarchy and allows us to calculate conceptual distance or similarity between predicted and reference metaphors even if its formulation is more abstract or differently phrased. By combining frame-based representations with contextualized embeddings, the study aims to evaluate metaphor understanding at the level of conceptual pattern recognition – a capability that bridges computational modelling and cognitive theories of metaphor (Lakoff and Johnson, 1980; 1999 – for an overview of different and more recent approaches see Dancygier & Sweetser, 2014 and Despot, 2024). The dataset's parallel nature further enables cross-lingual transfer experiments, where models trained on one language can be tested on another. To ensure that CroSloMet can serve as a shared resource for language technology research, supporting tasks such as frame mapping across languages and constructive evaluation of metaphor explanation models aligned with human interpretive criteria, CroSloMet is aligned with standardized semantic tagging schemes, and can be exported in interoperable formats providing benchmarks for corpus-based metrics.

This paper presents CroSloMet as both a robust dataset for metaphor research in under-resourced languages and an impetus for future method development in metaphor interpretation. By articulating a clear research agenda centered on structured annotations and nuanced evaluation, the work addresses persistent gaps in metaphor understanding with computational models. The proposed multi-level validation framework and cross-lingual experimental design position CroSloMet as a contribution to lexicography, computational linguistics, and cognitive semantics, making it relevant for figurative language modelling.

References

Chen, S., & Wang, Z. (2025). Challenges and Solutions for Large Language Models in Metaphor Translation of Political Texts. Open Journal of Applied Sciences, 15, 2360–2375. https://doi.org/10.4236/ojapps.2025.158158

Dancygier, B. & Sweetser, E. (2014). *Figurative language*. Cambridge University Press.

Despot, K. Š., Tonković, M., Brdar, M., Essert, M., Perak, B., Ostroški Anić, A., Nahod, B., & Pandžić, I. (2019). MetaNet.HR: Croatian Metaphor Repository. In M. Bolognesi, M. Brdar, & K. Š. Despot (Eds.), *Metaphor and Metonymy in the Digital Age* (pp. 123–146). John Benjamins Publishing Company. https://doi.org/10.1075/milcc.8.06des

Despot, K. Š., Ostroški Anić, A., & Veale, T. (2023). Somewhere along your pedigree, a bitch got over the wall! A proposal of implicitly offensive language typology. *Lodz Papers in Pragmatics*, 19(2), 385–414. https://doi.org/10.1515/lpp2023-0019

Despot, K. Š. (2024). *Metafora: Dekodiranje jezika imaginacije*. Institut za hrvatski jezik.

Despot, K. Š., Ostroški Anić, A., Gantar, P., Bon, M., Klemen, M., Robnik Šikonja, M., Krek, S., Perak, B., & Čibej, J. (2025). CroSloMet: A Structured Metaphor Dataset for Croatian and Slovene. *Fluminensia*, 37(2), 59–482. https://doi.org/10.31820/f.37.2.4

Devlin, J., Chang, M., 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. https://doi.org/10.18653/v1/N19-1423

Ge, M., Mao, R., Cambria, E. (2023). A survey on computational metaphor processing techniques: From identification, interpretation, generation to application. *Artificial Intelligence Review*, 56(S2), 1829–1895. https://doi.org/10.1007/s10462-023-10564-7

He, Q., Cheng, S., Li, Z., Xie, R., Xiao, Y. (2023). Can Pretrained Language Models Interpret Similes as Smart as Humans? a*rXiv*, abs/2203.08452. https://doi.org/10.48550/arXiv.2203.08452

Kim, J. C., Suh, S. G., Chilton, L. B., Xia, H. (2023). Metaphorian: Leveraging Large Language Models to Support Extended Metaphor Creation for Science Writing. In *Designing Interactive Systems Conference 23*. ACM. https://doi.org/10.1145/3563657.3595996

Klemen, M., Robnik-Šikonja, M. (2023). Neural Metaphor Detection for Slovene. In T. Erjavec & M. Eskevich (Eds.), *Selected papers from the CLARIN Annual Conference 2022* (pp. 77–89). Linköping Electronic Conference Proceedings 198. https://doi.org/10.3384/ecp198008

Lakoff, G. & Johnson, M. (1980). *Metaphors We Live by*. University of Chicago Press.

Lakoff, G. & Johnson, M. (1999). *Philosophy in the flesh*. Basic Books.

Liu, E., Cui, Ch., Zheng, K., Neubig G. (2022). Testing the Ability of Language Models to Interpret Figurative Language. In M. Carpuat, M.-C. de Marneffe, & I. V. Meza Ruiz (Eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4437–4452). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.naacl-main.330

Puraivan, E., Renau, I., & Riquelme, N. (2024). Metaphor identification and interpretation in corpora with ChatGPT. *SN Computer Science*, 5(8), 976. https://doi.org/10.1007/s42979-024-03331-0

Tong, X., Shutova, E., & Lewis, M. (2021). Recent advances in neural metaphor processing: A linguistic, cognitive and social perspective. In K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, & Y. Zhou (Eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4673–4686). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.372

Tong, X., Choenni, R., Lewis, M., & Shutova, E. (2024). Metaphor Understanding Challenge Dataset for LLMs. *ArXiv*, 2403.11810v1 [cs.CL]. https://doi.org/10.48550/arXiv.2403.11810

Yang, S., Zhang, D., Ren, J., Xu, Z., Zhang, X., Song, Y., Lin, H., & Xia, F. (2025). Cultural Bias Matters: A CrossCultural Benchmark Dataset and Sentiment-Enriched Model for Understanding Multimodal Metaphors. In W. Che, J. Nabende, E. Shutova, M. Taher Pilehvar (Eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics: Volume 1: Long Papers* (pp. 26301–26317). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.acl-long.1275

Contact information

Kristina Š. Despot
Institute for the Croatian Language
kdespot@ihjj.hr

Ana Ostroški Anić
Institute for the Croatian Language
aostrosk@ihjj.hr

Matej Klemen
University of Ljubljana
matej.klemen@fri.uni-lj.si

Emma Ferrett, Amir Ghorvei & Anastasia Riehl

A DICTIONARY FOR ALL CANADIANS

Establishing an Inclusive Standard for the Canadian English Dictionary

Keywords Canadian English; lexicography; standards; inclusivity

To create 'a dictionary for all Canadians' is one of the stated goals of the new Canadian English Dictionary (CED). But who are 'all' Canadians, and how can a single resource meet their varied needs? To help answer these questions and develop lexicographical principles and protocols, the team behind the CED is exploring the concept of an 'inclusive standard'. In this talk, we discuss our ongoing exploration of inclusivity and consider specific issues that arise in the drafting of entries.

Canada has been without an updated English dictionary for over twenty years (since *Canadian Oxford Dictionary*, 2nd ed., Barber, 2004). To address this gap, in 2022 a team of academics and editors launched the Society for Canadian English, a non-profit working group committed to creating a new resource. One goal is to produce an updated dictionary with changes to the language over the past twenty years; another important goal is to interrogate the 'standard' on which dictionaries are typically based and create a resource for 'all' that reflects the linguistic and cultural diversity of the country.

In Canada, the 'standard' variety of English typically characterizes the speech of urban, formally-educated people in the Central-West region from families with multiple generations in the country (Chambers, 1998). This conception of the standard comprises only 36% of the population (Dollinger, 2010) and excludes communities in the East and North, rural areas, those with less formal education, speakers of Indigenous Englishes, Anglophones in Québec, people in heritage language communities, and recent immigrants. A dictionary for 'all' must do better.

Our inclusivity goals are also necessarily decolonial. Dictionaries have historically been profoundly colonial in their construction and implementation, and projects like the CED must begin the work of practicing decolonial lexicographical methodologies. As well, an inclusive dictionary must consider sensitives around lexical items relating to specific social groups, including people with disabilities, racialized individuals and the LGBTQ+ community.

Representing the full range of Canadian English varieties and accents across different linguistic and social groups might achieve maximum inclusivity, but it could come at the cost of producing a resource with clear guidelines on how English is used in Canada, for those who seek such clarity. Dictionaries already struggle with a tension between descriptive and prescriptive aims and are used differently by editors, linguists, students and learners. An inclusivity goal makes balancing the needs and goals of diverse users even more challenging.

Given these issues, how do we define an 'inclusive standard', and how do we put our ideas into action in the dictionary-making process? After presenting a brief overview of the CED's history, methodological procedures and corpus research, we focus our discussion on inclusivity, providing examples in the realms of defining, pronunciations and etymology.

In the area of defining, one issue of consideration is our practice for the inclusion and presentation-order of variant spellings. In the case of *kayak*, for example, do we include the spelling *qajaq*, which reflects the word's Inuktitut roots, and if so, should it appear as the

headword or variant? Another issue is the treatment of 'offensive' terms. This includes our use of labels to clarify offensive historical terms (e.g., *half blood*, *paraplegic*) as well as our implementation of usage notes for terms that are historically (and in some contexts still) offensive, but that have been reclaimed as neutral or positive (e.g., *queer*, *Indian*).

Regarding pronunciations, the diversity of accents in Canada means that inclusivity considerations affect all aspects of the transcriptions. One issue is the choice of transcription system. We explain our decision to use the IPA — a divergence from North American tradition — to increase accessibility, especially for learners. Another issue is on whose speech we base the main pronunciation and, relatedly, how much variation we include. We discuss the challenges of choosing a 'model speaker' who is broadly representative of Canadian English as well as how to accommodate alternative pronunciations without embracing a level of detail better left to a dialect dictionary.

With etymologies, we have the opportunity to ground our inclusive standard in the decolonial practice of acknowledgment, similar to acknowledgments of unceded Indigenous land that settler-colonial peoples occupy. Many Canadian English words have an origin in an Indigenous language that may not be readily understood to the average speaker, for example, *kayak/qajaq*, *woodchuck*, and *squash* (vegetable). This etymological work highlights the debt of Canadian English to Indigenous languages of North America. It also acknowledges the view amongst Indigenous scholars such as Emma LaRocque (Cree and Métis) that "English is the new Native language" and that "it is English that is now serving to de-colonize and to unite" Indigenous Peoples (1990, 57).

In an inclusive dictionary, every Canadian should be able to open the resource and see their variety of English represented and treated with respect. We hope that our development of an inclusive standard will benefit not only the CED but serve as a model for future lexicographical projects.

References

Barber, K. (Ed.) (2004). *Canadian Oxford Dictionary* (2nd ed). Oxford University Press.

Chambers, J. (1998). English: Canadian varieties. In J. Edwards (Ed.), *Language in Canada* (pp. 252-72). Cambridge University Press.

Dollinger, S. (2010). *Creating Canadian English: The Professor, the Mountaineer, and a National Variety of English*. Cambridge University Press.

LaRocque, Emma. (1990). Preface or Here Are Our Voices–Who Will Hear? In H. Macfarlane and A.G. Ruffo (Eds.), *Introduction to Indigenous Literary Criticism in Canada* (pp. 47-60). Broadview Press.

Contact information

Emma Ferrett
Queen's University
emma.ferrett@queensu.ca

Amir Ghorvei
Queen's University
17mag@queensu.ca

Anastasia Riehl
Queen's University
riehla@queensu.ca

Emmanuel Sylvain Fomat

CAMFRANGLAIS IN THE AGE OF ARTIFICIAL INTELLIGENCE

Lexicographic Challenges, Opportunities and a Human-AI Approach to Documenting Linguistic Hybridity in Multilingual Cameroon

Keywords Camfranglais; human–AI collaboration; lexicography; low-resource languages; linguistic hybridity

In postcolonial societies, language often functions as both a site of power and resistance, a reality profoundly exemplified by Camfranglais in multilingual Cameroon. As Foucault (1978, pp. 95–96) argues, power inevitably generates resistance, and in Cameroon this resistance manifests linguistically through a hybrid sociolect that disrupts inherited colonial hierarchies. Emerging in the 1970s within a "communicative vacuum" (Kießling, 2015, p. 9) created by exoglossic language policies, Camfranglais combines French, English, Cameroonian Pidgin English, and numerous indigenous languages. Despite its widespread use and cultural significance, it remains substantially under-documented. Existing lexicographic resources resemble glossaries or linguistic guides constrained by static print formats, limited corpora, and the absence of systematic phonetic transcription. They also reveal a metalanguage mismatch: Kamdem Fonkoua (2015) provides English definitions with untranslated Camfranglais examples, creating barriers for Francophone users. Essome Bouti (2024) and Ndongo (2015) offer French metalanguage, which serves Francophone speakers but excludes Anglophone Cameroonians. Neither orientation serves speakers from the less-educated variety identified by Ebongue and Fonkoua (2010), who may have acquired French or English informally rather than through formal schooling. Consequently, a significant gap persists between the sociolinguistic reality of Camfranglais and its lexicographic representation.

The rise of artificial intelligence opens new possibilities for low-resource language documentation through large-scale data processing, pattern recognition, and phonetic prediction (Zhong et al., 2024; Wang, 2024; Midigo 2025). However, these opportunities are constrained by structural biases in current NLP systems, which are predominantly trained on high-resource and typologically stable languages. Ebrahimi et al. (2022, p. 6279) show that multilingual models achieve only 38.48% average zero-shot accuracy on genuinely low-resource languages. To address this tension, this paper proposes a Human–AI collaborative framework for documenting Camfranglais, grounded in the Function Theory of Lexicography (Bergenholtz & Tarp, 2002, 2003; Tarp, 2008a, 2008b). This framework informs the design of a bilingual, bidirectional digital dictionary with dual metalanguage access (French–Camfranglais and English–Camfranglais), intended to serve both Francophone and Anglophone users. Critical Discourse Analysis (CDA) is also applied to capture how Camfranglais expresses identity, ideology, and power relations in context (Fairclough, 1989; Blommaert & Bulcaen, 2000; Hidalgo Tenorio, 2011; Discourse Analyzer, 2024).

The study is further informed by an exploratory survey of forty Cameroonian speakers from diverse sociolinguistic backgrounds. Results confirm widespread informal use of Camfranglais and strong support for its documentation as a marker of national identity, while also revealing concerns that excessive standardisation could undermine its creativity and fluidity. In response, AI is positioned as a "copilot" subordinated to human expertise,

ensuring linguistic sovereignty and aligning with principles of anti-extractivism (Riofrancos, 2020).

Methodologically, the study relies on a deliberately constructed multi-source corpus capturing both historical written forms and contemporary oral-digital manifestations of Camfranglais. The written component includes the Grioo Forum (2004–2006), documenting early diasporic digital interaction, and the Bonaberi Forum (2008–2026), reflecting spontaneous written exchanges on everyday topics. The oral-digital component is drawn from the YouTube channel *Warman du Terre à Terre* (2013–2026). From this source, 1,583 video URLs were extracted to map lexical diversity, and 100 videos were fully transcribed to capture spontaneous speech, prosody, and emerging vocabulary. Together, these sources provide a heterogeneous corpus documenting the evolution of Camfranglais from street speech to digital discourse.

The analytical workflow operationalises a three-stage Human–AI feedback loop. First, computational tools identify candidate entries, frequency patterns, and collocations. Python was used for statistical extraction, while Sketch Engine (Kilgarriff et al., 2014) enabled deeper contextual analysis through concordances and Key Word In Context (KWIC) functions, particularly useful for identifying hybrid forms often missed by automated methods. Second, the Large Language Model Claude generated contextual definitions and etymological notes, while ChatGPT produced candidate phonetic transcriptions. Only standard French items were excluded; Cameroonian French forms with semantic or phonological deviation, as well as English items phonologically adapted by Camfranglophones, were retained as authentic lemmas. This procedure yielded approximately 2,850 lexical items, including multi-word expressions.

Finally, all outputs were validated by fourteen native Camfranglais speakers through independent five-point Likert-scale evaluations followed by consensus sessions. Results highlight both the strengths and limitations of AI-assisted lexicography. While AI accelerates extraction and contextual analysis, it struggles with tonal minimal pairs, pragmatic inversion, neologisms, hybrid morphology, and distinctions between Cameroonian and standard French. Nevertheless, validation achieved substantial inter-rater agreement (Fleiss' kappa = 0.74) (Landis & Koch, 1977), with high phonetic concordance: 94.2% perceptual agreement; 96.8% acoustic agreement using Praat (Boersma & Weenink, 2026).

This paper argues that a principled Human–AI approach is not merely a technical convenience but an epistemic necessity for documenting Camfranglais. By combining computational efficiency with native-speaker expertise, the study produces a pilot electronic dictionary that is phonologically informed, sociolinguistically grounded, and adaptable to ongoing lexical change. More broadly, it offers a transferable model for documenting hybrid urban varieties and demonstrates how artificial intelligence can support marginalized speech communities without reproducing extractive dynamics.

References

Bergenholtz, H. & Tarp, S. (2002). Die moderne lexikographische Funktionslehre. Diskussionsbeitrag zu neuen und alten Paradigmen, die Wörterbücher als Gebrauchsgegenstände verstehen. *Lexicographica*, 22, 145–155.

Bergenholtz, H. & Tarp, S. (2003). Two opposing theories: On H. E. Wiegand's recent discovery of lexicographic functions. *HERMES – Journal of Language and Communication in Business*, 16(31), 171–196. https://doi.org/10.7146/hjlcb.v16i31.25743

Blommaert, J. & Bulcaen, C. (2000). Critical discourse analysis. *Annual Review of Anthropology,* 29(1), 447–466. https://doi.org/10.1146/annurev.anthro.29.1.447

Boersma, P., & Weenink, D. (2026). *Praat: Doing phonetics by computer* (Version 6.4.65) [Computer program]. Retrieved April 4, 2026, from http://www.praat.org

Bonaberi Forum. (2026). *Bonaberi forum corpus* (2008–2026). https://www.bonaberi.com/forum/

Claude (Anthropic). (2026). *Claude* [Large language model]. https://claude.ai

Discourse Analyzer. (2024). *Societal structures in discourse analysis*. https://discourseanalyzer.com/societal-structures-in-discourse-analysis/

Ebongue, A. E., & Fonkoua, P. (2010). Le camfranglais ou les camfranglais? *Le Français en Afrique*, 25, 259–270.

Ebrahimi, A., Mager, M., Oncevay, A., Chaudhary, V., Chiruzzo, L., Fan, A., Ortega, J., Ramos, R., Rios, A., Meza Ruiz, I. V., Giménez-Lugo, G., Mager, E., Neubig, G., Palmer, A., Coto-Solano, R., Vu, T. & Kann, K. (2022). AmericasNLI: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (pp. 6279–6299). https://aclanthology.org/2022.acl-long.435

Fairclough, N. (1989). *Language and power*. Longman.

Foucault, M. (1978). *The history of sexuality* (Vol. 1). Pantheon Books.

Grioo Forum. (2006). *Kamers et sympathisants, parlons camfrangl comme au quat* (2004–2006). https://www.grioo.com/forum/viewtopic.php?t=974

Hidalgo Tenorio, E. (2011). Critical discourse analysis: An overview. *Nordic Journal of English Studies*, 10(1), 183–210. https://doi.org/10.35360/njes.247

Kamdem Fonkoua, H. (2015). *A dictionary of Camfranglais* (Vol. 107). Peter Lang. https://doi.org/10.3726/978-3-653-05363-0

Kießling, R. (2015). Hybrid juvenile discourse in Cameroon. In R. K. Agnihotri, C. Benthien & T. Oranskaia (Eds.), *Impure languages: Linguistic and literary hybridity in contemporary cultures* (pp. 155–178). Orient Black Swan.

OpenAI. (2026). *ChatGPT* [Large language model]. https://chat.openai.com

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography*, 1(1), 7–36. https://doi.org/10.1007/s40607-014-0009-9

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159–174. https://doi.org/10.2307/2529310

Midigo, J. O. (2025). AI-powered innovations for documenting and revitalizing African languages. *Cultural Arts Research and Development*, 5(2), 26–41. https://doi.org/10.55121/card.v5i2.517

Ndongo, V. (2016). *Dictionnaire du camfranglais*. Africa Stand Up.

Riofrancos, T. (2020). The origins of anti-extractivism. *Dissent Magazine*. https://dissentmagazine.org/online_articles/anti-extractive-politics/

Tarp, S. (2008a). *Lexicography in the borderland between knowledge and non-knowledge*. De Gruyter.

Tarp, S. (2008b). The third leg of two-legged lexicography. *Hermes – Journal of Language and Communication in Business*, 21(40), 117–131.

Wang, L. (2024). Artificial intelligence's role in the realm of endangered languages: Documentation and teaching. A*pplied and Computational Engineering*, 48, 123–129. https://doi.org/10.54254/2755-2721/48/20241249

Warman du Terre à Terre. (2026). *Warman du Terre à Terre* [YouTube channel] (2013–2026). https://www.youtube.com/@lewarman923

Zhong, T., Yang, Z., Liu, Z., et al. (2024). Opportunities and challenges of large language models for low-resource languages in humanities research. *arXiv*. https://doi.org/10.48550/arXiv.2412.04497

Contact information

Emmanuel Sylvain Fomat
University of Hildesheim
fomat@uni-hildesheim.de

Daniele Franceschi

GENDER REPRESENTATION AND EDITORIAL RESPONSIBILITY IN THE *OED*

A Case Study of *Female*, *Feminine*, *Male* and *Masculine*

Keywords Femininity; masculinity; gender stereotypes; *OED*

As a dictionary grounded in diachronic evidence, the *Oxford English Dictionary* (*OED*) occupies a central position in the tradition of historical lexicography, documenting not only the evolution of lexical meaning but also the cultural assumptions embedded in language over time. While the *OED* is explicitly descriptive and corpus-based, scholarship has shown that definitional framing and citation selection are shaped by editorial choices and socio-cultural ideologies (Baigent, Brewer & Larminie, 2005; Mugglestone, 2013; Norri, 2000; Vișan, 2021; Williams, 2023). Against this backdrop, this paper examines how gender is constructed and transmitted in the *OED* through a case study of the lemmas *female*, *feminine*, *male*, and *masculine*. The aim is to consider whether and to what extent gender representation has changed in the dictionary in light of the *OED*'s continuous revision process. More specifically, it assesses whether sustained editorial updating has substantially altered patterns of gender representation previously identified in lexicographic research.

These four lemmas were selected because of their apparent semantic neutrality and their foundational role in structuring gender categories in English, in contrast to more overtly evaluative terms such as *womanly*, *womanish*, *manly*, or *mannish*. The study adopts a qualitative, interpretive methodology, using present-day definitions as a point of semantic orientation while placing particular emphasis on the diachronic evidence provided by illustrative citations. All data are drawn from the *OED Online*, Third Edition (June 2025 update). Senses referring to human gender were analysed in full, while technical or grammatical uses were excluded unless their figurative extensions revealed ideological relevance. Particular attention is paid to citations from the nineteenth century onwards, a period of dense documentary coverage in the *OED* and of major ideological consolidation in relation to gender.

The analysis reveals persistent asymmetries in the representation of femininity and masculinity in the *OED*. Entries for *female* and *feminine* are more extensive and semantically elaborated than those for *male* and *masculine*, reflecting the marked status of femininity within the dictionary's treatment of gender. Across successive historical layers of citation, femininity is recurrently associated with emotional excess, passivity, superficiality, moral suspicion, and aesthetic or decorative qualities. Illustrative quotations frequently situate women in trivialised, domestic, or subordinate roles, or portray feminine agency as indirect or manipulative. Symbolic and metaphorical extensions—linking the feminine to flowers, planets, beauty products, or delicacy—further embed culturally inherited stereotypes within the historical record.

Masculinity, by contrast, is largely constructed as the unmarked and normative category. While *male* often functions as a default biological descriptor (as also in the case of *female*), *masculine* is consistently associated with strength, activity, authority, and intellectual or moral superiority. These qualities are rarely problematised in the historical citations and are frequently extended metaphorically to domains such as architecture, astrology, literature, and social organisation, reflecting long-standing models of hegemonic masculinity (Connell & Messerschmidt, 2005).

When masculine traits are attributed to women, they tend to be evaluated positively, whereas feminine traits applied to men are typically framed as deviations—an asymmetry well documented also in feminist linguistic scholarship (Kramarae, Treichler & Russo, 1985).

Although more recent citations in the *OED* begin to reflect changing conceptions of gender, including gender fluidity and non-binary identities, these developments remain unevenly distributed across the entries. Many senses of *male* and *masculine* continue to rely on relatively dated historical material, allowing older ideological models to dominate the semantic profile of these lemmas. As a result, the historical layering of citations becomes a mechanism through which outdated representations are preserved and reactivated in contemporary dictionary use.

The paper argues that historical dictionaries do not merely record linguistic change but actively mediate it through editorial selection and presentation (Pullum, 2018). By foregrounding the persistence of gendered asymmetries within a continuously revised dictionary, and by showing how definitions and citation practices can perpetuate gender hierarchies even in ostensibly neutral entries, this study contributes to current debates in historical lexicography and metalexicography. It concludes by suggesting that, unless definitions and citations are more carefully contextualised and balanced with contemporary evidence, even continuous editorial revision may do little to alter entrenched patterns of gender representation. A more reflexive approach would allow authoritative dictionaries such as the *OED* to preserve their historical record while mitigating the continued circulation of inherited gender ideologies.

References

Baigent, E., Brewer, C., & Larminie, V. (2005). Women and the archive: The representation of gender in the Dictionary of National Biography and the Oxford English Dictionary. *Archives*, 30, 13–35.

Connell, R. W., & Messerschmidt, J. W. (2005). Hegemonic masculinity: Rethinking the concept. *Gender & Society*, 19(6), 829–859. https://doi.org/10.1177/0891243205278639

Kramarae, C., Treichler, P. A., & Russo, A. (1985). *A feminist dictionary*. Pandora Press.

Mugglestone, L. (2013). Acts of representation: Writing the woman question in the Oxford English Dictionary. *Dictionaries*, 34, 39–65.

Norri, J. (2000). Labelling of derogatory words in some British and American dictionaries. *International Journal of Lexicography*, 13(2), 71–106. https://doi.org/10.1093/ijl/13.2.71

Oxford English Dictionary. (2025). *OED Online* (3rd ed., June 2025 update). Oxford University Press. https://www.oed.com

Pullum, G. K. (2018). Slurs and obscenities: Lexicography, semantics, and philosophy. In D. Sosa (Ed.), *Bad words: Philosophical perspectives on slurs* (pp. 168–192). Oxford University Press.

Vișan, R. (2021). On dictionaries and gender representations. *Gender Studies*, 20(1), 149–153. https://doi.org/10.2478/genst-2022-0010

Williams, D.-A. (2023). Women's words and the words of women in the Oxford English Dictionary. *Review of English Studies*, 74(316), 730–751. https://doi.org/10.1093/res/hgad052

Contact information

Daniele Franceschi
Roma Tre University
daniele.franceschi@uniroma3.it

Anne-Kathrin Gärtig-Bressan & Pia Carmela Lombardi

BILINGUAL DICTIONARIES WITH ITALIAN IN THE 19TH AND 20TH CENTURIES

The ALON Portal

Keywords Bilingual lexicography; ALON portal; German–Italian dictionaries; history of lexicography

The aim of this poster is to present ALON – *Archivio della Lessicografia dell'Otto-Novecento* (Archive of the Lexicography of the Nineteenth-Twentieth Century), an online portal dedicated to Italian lexicography. More specifically, this poster focuses on the section of the portal dedicated to bilingual dictionaries, particularly Italian–German ones.

During the 19th and 20th centuries, Italian lexicography underwent a process of renewal and reached its peak in terms of quantity, typology, and quality. This development also concerned bilingual dictionaries with Italian. The ALON project, funded by the European Union's Next Generation EU programme as an Italian PRIN project from 2023 to 2026, aimed at studying and documenting the dictionaries of this period. The online portal described above constitutes the core of the project and contains presentations and detailed entries on the single dictionaries. In addition to a section dedicated to one of the *Grandi Dizionari* of the time, the *Tommaseo-Bellini*, and another devoted to monolingual dictionaries, the portal also includes a section on bilingual dictionaries from the same period.

The state of research on Italian bilingual lexicography varies greatly depending on the different language pairs. For Italian–French and Italian–Spanish dictionaries, extensive repertoires already exist in Lillo (2020) and in the *Contrastiva* portal (directed by San Vicente). For English–Italian dictionaries, an important reference is O'Connor (1990) with a detailed bibliography of dictionaries until 1990. In the case of Italian–German lexicography, the most comprehensive work remains an unpublished thesis (Bruna 1983), alongside the overview by Bruna, Bray & Hausmann (1991) and single studies (e.g. Schweickard, 2000; Giacoma, 2012; *A scuola di Tedesco*, 2016; Gärtig, 2016). ALON initially focused on this language pair to fill a gap in research and to develop a model for the database structure that could later be extended to Italian bilingual dictionaries with other languages.

At the present time, the ALON portal includes descriptions of 20 Italian–German dictionaries in their first edition (48 when considering further editions), ranging from Domenico Antonio Filippi's *Dizionario italiano-tedesco e tedesco-italiano. Italienisch-deutsches und deutsch- italienisches Wörterbuch* (1817) to Martina Lucia Curcio's *Kontrastives Valenzwörterbuch der gesprochenen Sprache Italienisch-Deutsch* (1999). Among the dictionaries currently described on the platform, there are some works that represent milestones in Italian–German lexicography. One is Francesco Valentini's *Vollständiges italienisch-deutsches und deutsch-italienisches grammatisch-praktisches Wörterbuch. Gran Dizionario grammatico-pratico italiano-tedesco, tedesco-italiano* (1831–1836), one of the most important Italian–German bilingual dictionaries of the 19th century because of its many innovations, later adopted by other lexicographers. Another key work is Giuseppe Rigutini and Oskar Bulle's *Neues italienisch-deutsches und deutsch-italienisches Wörterbuch. Nuovo dizionario italiano-tedesco e tedesco-italiano* (1896–1900), which marked the first collaboration between an Italian and a German lexicographer. Giuseppe Rigutini was also one of the most important Italian lexicographers of the century and the author of several dictionaries. His collaboration with Oskar Bulle on an Italian–German dictionary therefore highlights the important role played by the two languages in relations between

the two nations. The platform also includes major dictionaries such as Henriette Michaelis' *Vollständiges Wörterbuch der italienischen und deutschen Sprache. Dizionario completo italiano-tedesco e tedesco-italiano* (1879–1881), as well as so-called "pocket editions" such as Friedrich Ernst Feller's *Nuovo dizionario portatile italiano tedesco, tedesco Italiano. Neuestes Taschen-Wörterbuch der italienischen und deutschen Sprache* (1851).

The database has a modular structure and includes both extensive texts and synthetic information on the dictionaries described. The level of detail varies according to the type of work considered. In addition to identification data for the dictionaries and an overview of their most important sources and research works on them, longer texts may describe their history (conception, publication history, reception, and circulation), structure, and content (paratexts, macro- and microstructure). Under the section "Materials and Resources," links to digitized full texts are provided where available, together with documents such as images or archive material. This is followed by the bibliography. For bilingual dictionaries, the portal also provides additional summary information, including a transcription of the title page, a reconstructed table of contents, information on entry structure (besides the above-mentioned works cfr. Hausmann, 1989; Marello, 1989; Marello & Rovere, 1999), sources, distribution in Italy, and details of the consulted library. The portal takes into account not only the first edition of a dictionary, but also later editions, which may differ significantly.

The names of the dictionary authors are linked to a separate section describing their lives and works. This allows to illustrate the connections between lexicographical projects: for example, a person may appear as a reviser in one dictionary project and as an author in another.

All dictionary descriptions in the portal are assigned a DOI (Digital Object Identifier). The database is conceived as an expandable resource and offers external authors the opportunity to publish, after peer review, their research on Italian dictionaries from the 19th and 20th centuries in the form described above.

References

A scuola di tedesco: censimento sistematico della manualistica per l'insegnamento e l'apprendimento del tedesco nelle biblioteche trentine (1511–1924). Schede a cura di Manuela Rizzoli, direzione scientifica di Paola Maria Filippi. Provincia autonoma di Trento, Soprintendenza per i Beni culturali, Ufficio per i Beni archivistici, librari e Archivio, 2016

ALON: *Archivio della Lessicografia dell'Otto-Novecento.* Retrieved May 18, 2026, from https://archivio-alon.it/

Bruna, M. L. (1983). *La lessicografia italo-tedesca* [unpublished thesis, Università degli Studi di Udine].

Bruna, M. L., Bray, L. & Hausmann, F. J. (1991). Die zweisprachige Lexikographie Deutsch-Italienisch, Italienisch-Deutsch. In F. J. Hausmann et al. (Eds.), *Wörterbücher. Dictionaries. Dictionnaires. Ein internationales Handbuch zur Lexikographie. An International Encyclopedia of Lexicography. Encyclopédie internationale de lexicographie* (vol. 3, 3013–3019). De Gruyter.

Contrastiva: Portal de lingüística contrastiva español-italiano. Retrieved May 18, 2026, from https://www.contrastiva.it/wp/

Curcio, M. L. (1999). *Kontrastives Valenzwörterbuch der gesprochenen Sprache Italienisch-Deutsch. Grundlagen und Auswertung.* IDS.

Feller, F. E. (1851). *Nuovo dizionario portatile italiano tedesco, tedesco italiano. Arricchito d'una gran quantità di vocaboli relativi al commercio, alle strade ferrate e ai vapori. Neuestes Taschen-Wörterbuch der italienischen und deutschen Sprache für Reisende und zum Schulgebrauch. Mit Berücksichtigung der gesammten kaufmännischen Terminologie, sowie der neuesten Ausdrücke in Bezug auf Eisenbahnen, Dampfböte &c.* Teubner.

Filippi, D. A. (1817). *Dizionario italiano-tedesco e tedesco-italiano. Italienisch-deutsches und deutsch-italienisches Wörterbuch*. Heubner e Volke & Carl Cnobloch.

Gärtig, A.-K. (2016). *Deutsch-italienische Lexikographie vor 1900. Die Arbeiten des Sprach- und Kulturmittlers Francesco Valentini (1789-1862)*. De Gruyter.

Giacoma, L. (2012). Oltre 100 anni di evoluzione dei dizionari di Tedesco-Italiano. Analisi comparativa della voce *cadere*. In L. Cinato et al. (Eds.), *Intrecci di lingua e cultura. Studi in onore di Sandra Bosco Coletsos* (pp. 149–170). Aracne.

Hausmann, F. J. (1989). Die Markierung im allgemeinen einsprachigen Wörterbuch: eine Übersicht. In F. J. Hausmann et al. (Eds.), *Wörterbücher. Dictionaries. Dictionnaires. Ein internationales Handbuch zur Lexikographie. An International Encyclopedia of Lexicography. Encyclopédie internationale de lexicographie* (vol. 1, pp. 649–657). De Gruyter.

Lillo, J. (Ed.) (2020). *1583-2010: Quattro secoli e più di lessicografia italo-francese. Repertorio analitico di dizionari bilingue*. Clueb.

Marello, C. (1989). *Dizionari bilingui con schede sui dizionari italiani per francese, inglese, spagnolo, tedesco*. Zanichelli.

Marello, C. & Rovere, G. (1999). Mikrostrukturen in zweisprachigen Wörterbüchern Deutsch- Italienisch/ Italienisch-Deutsch. In H. E. Wiegand (Ed.), *Studien zur zweisprachigen Lexikographie mit Deutsch* 4 (pp. 177–206). Olms.

Michaelis, H. (1879–1881). *Vollständiges Wörterbuch der italienischen und deutschen Sprache mit besonderer Berücksichtigung der technischen Ausdrücke des Handels, der Gewerbe, der Wissenschaften, des Kriegs- und Seewesens, der Politik u.s.w. Dizionario completo italiano-tedesco e tedesco-italiano con riguardo speciale alle espressioni tecniche del commercio, delle scienze, dell'industria, della guerra e della marina, della politica ecc.* Brockhaus.

O'Connor, D. (1990). *A History of Italian and English Bilingual Dictionaries*. Leo S. Olschki.

Rigutini, G. & Bulle, O. (1896–1900). *Neues italienisch-deutsches und deutsch-italienisches Wörterbuch. Nuovo dizionario italiano-tedesco e tedesco-italiano*. Tauchnitz & Hoepli.

Rizzoli, M. (2016). *A scuola di tedesco: Censimento sistematico della manualistica per l'insegnamento e l'apprendimento del tedesco nelle biblioteche trentine (1511–1924)*. Soprintendenza per i Beni culturali, Ufficio per i Beni archivistici, librari e Archivio.

Schweickard, W. (2000). Zur zweisprachigen Lexikographie Deutsch und Italienisch. In H. E. Wiegand (Ed.), *Studien zur zweisprachigen Lexikographie mit Deutsch* 5 (pp. 71–86). Olms.

Valentini, F. (1831–1836). *Vollständiges italienisch-deutsches und deutsch-italienisches grammatisch-praktisches Wörterbuch nach den neuesten und besten Quellen beider Sprachen bearbeitet und mit ungefähr 40,000 technischen und wissenschaftlichen Wörtern und Ausdrücken und beinahe 60,000 neuen Artikeln versehen. Gran Dizionario grammatico-pratico italiano-tedesco, tedesco-italiano, composto sui migliori e più recenti vocabolarii delle due lingue, ed arricchito di circa 40,000 voci, e termini proprii della scienza ed arti, e di 60,000 nuovi articoli*. Johann Ambrosius Barth.

Contact information

Anne-Kathrin Gärtig-Bressan
Università degli Studi di Trieste
akgaertig@units.it

Pia Carmela Lombardi
Università degli Studi di Trieste
piacarmela.lombardi@units.it

Marius Glebus

CHOOSING OPTIMAL AI PROMPT INSTRUCTIONS FOR NEOLOGISM DETECTION

Keywords Neology; prompt engineering; generative artificial intelligence

The past few years have seen numerous attempts to make use of generative artificial intelligence (AI) in the field of lexicography. One such field is neology – the detection of new words and the preparation of their lexicographic entries for further study (cf. Zheng et al., 2024). When identifying neologisms in large collections of texts with the help of AI, existing neologism data – usage examples already collected by lexicographers – can be used for two purposes: (1) as positive examples included in AI prompts and (2) as material for evaluating model performance (cf. Tang et al., 2025). Since the number of neologisms in a dataset of usage examples is known, it is possible to calculate not only precision but also recall, enabling an objective evaluation of model performance.

Traditional methods of collecting neologisms are still very effective. The *Database of Lithuanian Neologisms* (DLN) relies on two traditional collection methods: editors regularly record neologisms encountered while reading news portals and social media; members of the general public submit neologism candidates they have identified via a form on the DLN website. As of 2026, the DLN contains more than 12,000 detailed neologism entries with definitions and usage examples. Because individual entries may have several examples, the database currently includes more than 43,000 usage examples in total. Many of these examples consist of multiple sentences, thus making the DLN a source for a very neologism-rich corpus comprising several million words. An additional advantage of manual collection is that neologisms can be documented from sources that are difficult to access for automated tools: social media posts, comments on YouTube and news portals, radio and TV recordings.

A dataset of this size and quality provides a strong basis for developing AI tools for automated neologism extraction from web texts and definition generation.

This presentation discusses prompt engineering experiments aimed at identifying an optimal prompt for the task of neologism detection, incorporating usage examples from the DLN. The experiments focus on the component of the prompt known as the instruction (or task description). Prompt engineering guidelines (Liu et al., 2024) typically describe a prompt as consisting of several elements: (1) the instruction, which tells the model what task to perform; (2) a role specification defining the model's assumed role; (3) examples with corresponding correct responses; (4) formatting, style and length requirements for the output; (5) the input material that the model is asked to process.

Two alternative instruction strategies for neologism detection were tested. The first strategy instructs the model to process a given text and return only the neologisms it identifies (*extraction-style*). The second strategy instructs the model to examine each word in the text individually and determine whether it constitutes a neologism (*classification-style*). Using an open-weight language model, a series of experiments was conducted to evaluate the performance of both prompting strategies. Although the difference between these formulations may appear relatively minor, it results in substantial differences both in output quality and in the computational resources required to complete the task.

The experiments demonstrate that both prompting strategies achieve satisfactory performance in neologism detection. The classification-style prompt shows a slight advantage in the total number of correctly identified neologisms, whereas the extraction-style prompt demonstrates a significant advantage in terms of lower noise, i.e. fewer words incorrectly identified as neologisms. An even greater difference between the two strategies emerges from the perspective of practical application. When processing the same text on identical hardware, the extraction-style prompt required nearly twenty times less processing time than the classification-style prompt.

Both prompting strategies exhibit similar weaknesses. Despite explicit instructions, the model may return words containing introduced misspellings, provide grammatically altered forms instead of the original forms occurring in the analyzed text, omit relevant words from the output, reorder words relative to their order in the source text, or introduce words that do not appear in the analyzed text at all (hallucinations).

Further research should include a lexicographic evaluation of how the two prompting strategies perform in detecting specific categories of neologisms, such as derivatives, compounds, and borrowings. From a technical perspective, methods for mitigating model-induced distortions of analyzed words should also be explored.

The experiments indicate that current large language models can provide substantial assistance to lexicographers in the collection of neologisms. Nevertheless, professional linguistic expertise remains necessary for the final evaluation of candidate neologisms and the removal of false positives.

References

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2024). Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics,* 12, 157–173. https://doi.org/10.1162/tacl_a_00638

Tang, Y., Tuncel, D., Körner, Ch., & Runkler, T. (2025). The Few-shot Dilemma: Over-prompting Large Language Models. *Proceedings the of 3rd International Conference on Foundation and Large Language Models (FLLM)* (pp. 134–141).

Zheng, J., Ritter, A., & Xu, W. (2024). NEO-BENCH: Evaluating Robustness of Large Language Models with Neologisms. *Proceedings of the 62nd Annual Meeting of the ACL (Volume 1: Long Papers)* (pp. 13885–13906).

Contact information

Marius Glebus
Institute of the Lithuanian Language
marius.glebus@lki.lt

Wanjing Han & Gilles-Maurice de Schryver

CAN LLMS CAPTURE CHINA'S EMERGING ECONOMIC NEOCONCEPTS LEXICOGRAPHICALLY?

Keywords GenAI; Chinese neologisms; Chinese–English LSP dictionaries; economic discourse; reception-oriented

Recent years have seen a rapid integration of Generative Artificial Intelligence (GenAI) into lexicographic practice. Studies have demonstrated that Large Language Models (LLMs) are increasingly capable of generating dictionary definitions, example sentences, and structured lexicographic entries (de Schryver, 2023, 2024, 2025a, 2025b; Lew, 2023; Lew et al., 2024; Chen et al., 2025). At the same time, user research indicates that traditional paper and digital dictionaries are no longer perceived as sufficient for many of their intended real-world tasks, especially for Chinese (Han & de Schryver, 2026).

Although a growing body of work has examined LLM performance for culture-specific tasks in general (Li et al., 2024), and culture-specific lexicographic tasks involving Chinese in particular (Han & de Schryver, 2025; Wu & Wang, 2026), existing research has predominantly focused on relatively stable and general-language domains (but see Barik & Ed-Dali, 2025, for Shakespearean lexicography, or Heinisch, 2025, for multilingual terminology). Little attention has been given to contemporary Chinese economic and socio-commercial neologisms, especially those emerging from platform economies and informal market practices, where meanings are institutionally embedded, culturally conditioned, and often semantically fluid. Such items present distinct challenges for lexicographic modelling, as they require not only semantic accuracy but also cultural framing. Moreover, comparative evaluations of LLMs in lexicographic contexts have largely centred on a limited range of Generative Pre-trained Transformer (GPT) models, with relatively little systematic discussion of potential differences between for instance US-developed and China-developed GPT models in handling cultural-specific lexical material. As Chinese neologisms are frequently rooted in local regulatory frameworks, economic policies, and technology-based market dynamics, the extent to which different LLMs interpret and structure such content remains insufficiently explored.

To address these issues, the present study investigates the generation of user-oriented lexicographic answers for contemporary Chinese economic and socio-commercial neologisms across four LLMs. The models tested were GPT-5.5 in the High setting (OpenAI, 2026), Gemini 3.5 Flash (Google, 2026), DeepSeek-V4-pro (DeepSeek, 2026), and Qwen3.7-Plus (Qwen, 2026). These models were selected to compare two US-based systems with two China-based systems. We selected Chinese economic neoconcepts from the '2024 Media Neologism List', published as part of 'China's Language Life Reports' in October 2025. Based on their semantics, these neologisms were manually categorised into six domains: (i) policy instruments, (ii) financial mechanisms, (iii) strategic industries, (iv) platform and consumption micro-economies, (v) informal practices, and (vi) geoeconomic discourses. Among these categories, terms associated with the fifth category, namely informal practices, were identified as particularly complex due to their cultural embeddedness, institutional specificity, and semantic variability. We therefore selected five representative economic neoconcepts from this domain. A detailed analysis was then performed, not by a machine (as for instance done in Zou & Yu, 2025), but by humans, to assess which of the four LLM models gives the best overall result. Observe that we purposely use two American versus two Chinese models, not only to evaluate which model is best, but also to check whether the claims by OpenAI (see Financial Times, 2025) and

Anthropic (see Anthropic, 2026) are correct when they say that Chinese LLM models like DeepSeek are mere copycats that 'distilled' their chatbots, "feeding them prompts in order to learn from and mimic their responses" (The Economist, 2026). Until recently, "The more a model read, the cleverer it became" (The Economist, 2026), and this is still the case for the big American frontier models today; whereas Chinese models would merely produce 'synthetic data'. If true, one would assume that highly complex Chinese economic and socio-commercial neoconcepts have no chance to be rendered better by Chinese chatbots . . . and yet, our results seem to indicate the exact opposite!

Our study thus seeks to contribute to a better understanding of how GenAI systems represent culturally-embedded lexical knowledge in a Chinese–English lexicographic Language for Specific Purposes (LSP) setting. More specifically, our findings highlight the potential of LLM-based methods for compiling reception-oriented Chinese–English LSP neologism dictionaries.

References

Anthropic (2026). *Detecting and preventing distillation attacks. Anthropic – Announcements,* 23 February 2026. https://www.anthropic.com/news/detecting-and-preventing-distillation-attacks

Barik, A. and Ed-Dali, R. (2025). The representation of Shakespearean corpus in ChatGPT's repertoire. *Revue Etudes Langue et Littérature* 4(2), 63–79.

Chen, X., Pan, Y. & Yang, Y. (2025). 外向型汉语学习词典编纂的研究——以 DeepSeek 为例 [A preliminary study of the large language model-assisted compilation of learner-oriented Chinese dictionaries: A case study of DeepSeek]. *云南师范大学学报 (对外汉语教学与研究版) [Journal of Yunnan Normal University (Teaching and Researching on Chinese as a Foreign Language)]* 23(2), 11–23.

DeepSeek (2026). *DeepSeek-V4-Pro* [Large language model]. https://chat.deepseek.com/

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. (2024). The road towards fine-tuned LLMs for lexicography. In S. Krek (Ed.), *Book of Abstracts of the Workshop "Large Language Models and Lexicography"*, 8 October 2024, Cavtat, Croatia (pp. 6–11). ELEXIS Association.

de Schryver, G.-M. (2025a). Out-of-the-box GPTs for lexicography. *Lexikos* 35(2), 362–396.

de Schryver, G.-M. (2025b). Customised GPTs for lexicography. *Lexikos* 35(2), 397–422.

Financial Times (2025). OpenAI says it has evidence China's DeepSeek used its model to train competitor. White House AI tsar David Sacks raises possibility of alleged intellectual property theft. *Financial Times – Artificial Intelligence,* 29 January 2025. https://www.ft.com/content/a0dfedd1-5255-4fa9-8ccc-1fe01de87ea6

Google (2026). *Gemini 3.5 Flash* [Large language model]. https://gemini.google.com/

Han, W. & de Schryver, G.-M. (2025). From user feedback on Chinese-English translation and dictionary tools, to DeepSeek as the ultimate translator and lexicographer. In Y. Gao & X. Zhang (Eds.), *Proceedings of the 18th International Conference of the Asian Association for Lexicography: Traditionality and Creativity in the Era of AI* (pp. 53–62). Fudan University.

Han, W. & de Schryver, G.-M. (2026). Perceptions of language service technologies and reference works by Chinese users in the era of Generative AI. *International Journal of Lexicography* 39, 1–23.

Heinisch, B. (2025). Next-gen terminology: Transforming terminology work with large language models. *Across Languages and Cultures* 26(S), 64–80.

Lew, R. (2023). ChatGPT as a COBUILD lexicographer. *Humanities and Social Sciences Communications* 10(704), 1–10.

Lew, R., Ptasznik, B. & Wolfer, S. (2024). The effectiveness of ChatGPT as a lexical tool for English, compared with a bilingual dictionary and a monolingual learner's dictionary. *Humanities and Social Sciences Communications* 11(1324), 1–10.

Li, C., Chen, M., Wang, J., Sitaram, S. & Xie, X. (2024). CultureLLM: Incorporating cultural differences into Large Language Models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak & C. Zhang (Eds.), *Proceedings of the 38th Conference on Neural Information Processing Systems* (pp. 84799–84838). Vancouver: Curran Associates.

OpenAI (2026). *GPT-5.5 in the High setting* [Large language model]. https://chatgpt.com/

Qwen (2026). *Qwen3.7-Plus* [Large language model]. https://chat.qwen.ai/

The Economist (2026). American labs say China's AI tigers are copycats. DeepSeek's new model has American officials and firms on edge. *The Economist,* 28 February 2026. https://www.economist.com/china/2026/02/25/american-labs-say-chinas-ai-tigers-are-copycats

Wu, M. & Wang, D. (2026). Automatic compilation of a Pre-Qin philosophy lexicon via large language models. *npj Heritage Science* 14, 1–11.

Zou, Y. & Yu, S. (2025). Can Large Language Models understand Chinese neologisms? In L. Wang, R. Tong, S.F.S. Juan, Y. Lu, P.P. Tan, S. Saee & M. Dong (Eds.), *Proceedings of the 29th International Conference on Asian Language Processing* (pp. 13–18). Institute of Electrical and Electronics Engineers.

Contact information

Wanjing Han
Ghent University
wanjing.han@ugent.be

Gilles-Maurice de Schryver
Ghent University & Universty of Pretoria
gillesmaurice.deschryver@ugent.be

Wanjing Han & Gilles-Maurice de Schryver

CORPUS-BASED AND AI-ASSISTED DESIGN OF A CHINESE–ENGLISH BUSINESS DICTIONARY

Keywords Artificial intelligence (AI); generative AI (GenAI); DeepSeek; Chinese–English parallel corpus; Chinese–English terminology dictionary; cultural sensitivity

Recent research in lexicography has shown that the integration of generative AI into dictionary-making can work well for bilingual tasks (Chen, 2025; de Schryver, 2025; Han & de Schryver, 2025). The present research reports on a corpus-based study of Chinese–English parallel business news and explores its implications for bilingual business lexicography in the age of AI. The study is based on a purpose-built Chinese–English parallel corpus drawn from the "Bilingual News" section of *China Daily*, comprising over 36,000 Chinese tokens and 42,000 English tokens. Although the corpus is relatively small and restricted to a single source, it includes all available bilingual business news published by this major official outlet in 2024. It can therefore be seen as a complete dataset within a clearly defined domain.

The English texts are typically original reports aimed at international audiences, while the Chinese texts are translations for domestic readers. Using the Sketch Engine, the analysis combines single-word keyword extraction and multi-word term extraction, with comparisons against large general-news reference corpora in both languages. A minimum frequency threshold of 5 is applied, and items with very low keyness scores are excluded, as they do not show domain-specific salience. This approach makes it possible to identify lexically salient items that are characteristic of contemporary business and business-related policy discourse.

The findings show clear cross-linguistic asymmetries. Chinese business texts tend to be data-driven and domestically oriented, favouring compact noun compounds, numerical indicators, and institution-specific terminology that compress complex economic relations into dense nominal structures. English texts, by contrast, rely more heavily on abstract policy framing, analytic syntactic constructions, and standardised acronyms and compound expressions, with repeated use of established policy phrases. These asymmetries should be interpreted with caution. As the English texts are produced within Chinese official media contexts, they may reflect patterns of China English and may differ from the expressions typically used in media from English-speaking countries, rather than purely language-internal variation.

Across both languages, multi-word units show markedly higher keyness values than single-word items. This suggests that meaning in this domain is primarily conveyed through fixed expressions, collocations, and policy phrases. While some high-frequency Chinese–English multi-word pairs have stabilised as standard equivalents, many Chinese business terms do not have direct lexical counterparts in English. These items are often culture- or institution-specific and therefore call for a paraphrastic or definition-based treatment rather than a word-for-word translation.

Building on these findings, the study proposes several lexicographic implications. First, it argues for prioritising multi-word units as dictionary headwords, as these units capture recurrent usage and encode domain-specific concepts more directly. Second, it supports a morpheme-plus-collocation entry structure. Under this structure, core lexical elements are systematically linked to their most frequent and productive multi-word patterns. Third, it advances a dual-track dictionary model. This model is designed to accommodate both the

production-oriented needs of Chinese users and the reception-oriented needs of foreign users, allowing lexicographic content to be organised according to different user situations and functions.

To explore the role of GenAI in contemporary lexicographic practice, the study uses an AI chatbot, DeepSeek-V3, to generate preliminary dictionary entries. These are generated entirely by GenAI, with prompts iteratively refined to improve the quality and structure of the output. The prompt is shown in Addendum A. DeepSeek-V3 was provided with compilation notes, as seen in Addendum B. Examples of generated entries are shown in Addenda C and D. To provide a point of comparison for the material seen in Addenda C and D, Addendum E shows the official English translation and explanation of the same Chinese term, namely "新质生产力" (New Quality Productive Forces).

During the talk, the study will also compare the AI-generated entries with human-produced lexical resources. When accessed via dictionary apps, reference works such as the *New Century Chinese–English/English–Chinese Dictionary* (2026) and the *Youdao Dictionary* (2026) mainly provide equivalents, but do not offer full definitions nor usage examples. Overall, then, the present study demonstrates how corpus evidence and GenAI chatbots can jointly inform the next-generation of bilingual specialised dictionary design. By integrating corpus-based insights and AI-assisted design, the paper contributes to ongoing discussions on dictionary-making processes in the age of GenAI.

References

Chen, Y. (2025). A Comparative study on the effectiveness of AI chatbots and dictionary apps for lexical tasks and retention. *Lexikos*, 35(1), 157–182.

de Schryver, G.-M. (2025). Customised GPTs for lexicography. *Lexikos*, 35(2), 397–422 + twenty-nine pages of supplementary material online.

Han, W. & de Schryver, G.-M. (2025). From user feedback on Chinese–English translation and dictionary tools, to DeepSeek as the ultimate translator and lexicographer. In Y. Gao & X. Zhang (Eds.), *AsiaLex 2025, September 11-13, Fudan, Shanghai, Proceedings of the 18th International Conference of the Asian Association for Lexicography: Traditionality and Creativity in the Era of AI* (pp. 53–62). Fudan University.

NetEase. (2026). *有道词典* [Youdao Dictionary] (Version 11.1.12) [iOS].

Shanghai Haidi Digital Publishing Technology Co., Ltd. (2026). *新世纪英汉汉英大词典* [New Century English–Chinese/Chinese–English Dictionary] (Version 3.0.6) [iOS].

Contact information

Wanjing Han
Ghent University
wanjing.han@ugent.be

Gilles-Maurice de Schryver
Ghent University & Universty of Pretoria
gillesmaurice.deschryver@ugent.be

Addendum A: The prompt given to DeepSeek-V3

我在编纂一本词典，语言方向是从中文到英文，请帮我给"新质生产力"(New Quality Productive Forces) 设计dual track 词条，用户分别是Chinese Users (L1), foreign users (L2)。

English translation:
I am compiling a dictionary. The language direction is from Chinese to English. Please help me design a dual-track dictionary entry for "新质生产力" (New Quality Productive Forces). The target users are Chinese users (L1) and foreign users (L2).

Addendum B: Compilation notes for entry design by DeepSeek - V3

编纂说明 | Compilation Notes:

1. **双轨设计：** L1轨侧重为中文用户提供精准的英文对应、深层阐释及语境，服务于翻译与对外阐释；L2轨侧重为外语用户厘清概念内涵、背景及用法，服务于理解当代中国话语。
2. **定义策略：** L1定义更注重与国内政治经济学话语体系衔接；L2定义则从功能与特征入手，更符合外语用户的认知习惯。
3. **术语关联：** 均提供核心相关术语，构建微型概念网络，增强实用性。

此设计旨在确保词条在学术准确性、政策敏感性和跨文化传播有效性之间取得平衡。

English translation:

Compilation Notes:

1. **Dual-track design**: The L1 track focuses on providing Chinese users with precise English equivalents, deeper explanations, and usage guidance, serving translation and external communication needs. The L2 track aims to help foreign users understand the concept, background, and usage, supporting the comprehension of contemporary Chinese discourse.
2. **Definition strategy**: L1 definitions place more emphasis on alignment with domestic political and economic discourse systems. L2 definitions start from functions and features, and are more in line with the cognitive habits of foreign users.
3. **Terminology relations**: Both tracks provide core related terms and construct a micro-level conceptual network, which improves practical usability.

This design aims to achieve a balance between academic accuracy, policy sensitivity, and effectiveness in cross-cultural communication.

Addendum C: Dictionary entry for “New Quality Productive Forces” targeted at Chinese users (L1)

Track 1: For Chinese Users (L1) | 面向中文母语使用者

词条：新质生产力

拼音： xīn zhì shēngchǎn lì

核心定义：
指在科技创新驱动下，由技术革命性突破、生产要素创新性配置、产业深度转型升级而催生的先进生产力质态。其核心特征是高科技、高效能、高质量，摆脱传统经济增长方式和生产力发展路径，代表生产力的跃迁。

英文对应译名：
New Quality Productive Forces (主流译法)
亦作：New-quality Productive Forces, New Quality Productivity

深度阐释：
“新质生产力”是中国特色社会主义政治经济学的最新重要概念，强调以**创新**（特别是科技创新）为主导，**劳动者、劳动资料、劳动对象**及其优化组合的跃升。其载体是**战略性新兴产业**（如人工智能、新能源、新材料）和**未来产业**。它区别于主要通过增加资源投入（劳动力、资本）驱动的传统生产力，更注重全要素生产率的提升、数字智能与绿色发展的深度融合。

语境应用：

- **宏观政策语境：** 加快发展新质生产力，是推动高质量发展的内在要求和重要着力点。Accelerating the development of New Quality Productive Forces is essential for promoting high-quality development.
- **产业/企业语境：** 该公司通过布局人形机器人赛道，积极培育新质生产力。The company is actively fostering New Quality Productive Forces by venturing into the humanoid robotics sector.
- **科技语境：** 原创性、颠覆性科技创新是形成新质生产力的关键。Original and disruptive technological innovation is key to forming New Quality Productive Forces.

相关术语：

- **高质量发展：** high-quality development
- **战略性新兴产业：** strategic emerging industries
- **未来产业：** future industries
- **全要素生产率：** total factor productivity (TFP)
- **创新驱动发展：** innovation-driven development

Addendum D: Dictionary entry for “New Quality Productive Forces” targeted at foreign users (L2)

Track 2: For Foreign Users (L2) | 面向外语使用者

Entry: New Quality Productive Forces

Chinese Original: 新质生产力 (xīn zhì shēngchǎn lì)

Core Definition:
A Chinese political-economic concept describing a new, advanced form of productive power. It arises from revolutionary technological breakthroughs, innovative allocation of production factors (labor, capital, data, etc.), and deep industrial transformation. It is characterized by **high technology, superior efficiency, and high quality**, moving beyond traditional growth models that rely on sheer input expansion.

Key Conceptual Breakdown:

- **New (新):** Not incremental, but a qualitative leap in how production is organized and value is created.
- **Quality (质):** Emphasizes intrinsic superiority, sustainability, and value-added output.
- **Productive Forces (生产力):** The combined capacity to produce goods and services (from Marxist political economy).

What It Entails:

1. **Driven by:** Cutting-edge innovation, especially in science and core technologies.
2. **Embodied in:** Strategic emerging industries (AI, new energy, biotechnology) and future industries.
3. **Aims for:** Enhancing **total factor productivity**, integrating digitalization and greening, and achieving high-quality development.

Usage Notes:

- This is a **high-level policy term**, frequently used in official discourse, development plans, and economic analysis of contemporary China.
- It signals a shift from scale-and-speed growth to innovation-and-sustainability-led growth.
- Often appears in contexts discussing China's technological self-reliance and industrial modernization.

Example Sentences:

- China is focusing on developing **New Quality Productive Forces** to power its next phase of economic growth.
- The government report emphasized fostering **New Quality Productive Forces** through support for research in artificial intelligence and clean energy.
- Cultivating **New Quality Productive Forces** is seen as essential for gaining an edge in global competition.

Related Terms:

- **High-quality development:** 高质量发展
- **Strategic emerging industries:** 战略性新兴产业
- **Innovation-driven development:** 创新驱动发展
- **Scientific and technological self-reliance:** 科技自立自强

Addendum E: Official English translation and explanation of "新质生产力"

(https://wb.beijing.gov.cn/home/wswm/yyhj/fyyd/202405/t20240521_3689647.html)

新质生产力

New quality productive forces

是由技术革命性突破、生产要素创新性配置、产业深度转型升级而催生的当代先进生产力，以劳动者、劳动资料、劳动对象及其优化组合的质变为基本内涵，以全要素生产率提升为核心标志。2023年9月习近平总书记在黑龙江考察调研期间提出，整合科技创新资源，引领发展战略性新兴产业和未来产业，加快形成新质生产力。习近平总书记在中央经济工作会议上进一步强调，要以科技创新推动产业创新，特别是以颠覆性技术和前沿技术催生新产业、新模式、新动能，发展新质生产力。

They refer to advanced productive forces emerging from revolutionary technological breakthroughs, innovative reallocation of production factors, and deep industrial transformation and upgrading. It is essentially about qualitative changes in labor force, means of production, objectives of labor and the combination thereof, as represented by a substantial increase in total factor productivity.

The concept was put forward by General Secretary Xi Jinping during his inspection tour to Heilongjiang Province in September 2023. He emphasized the necessity of integrating technological innovation resources to fuel the development of emerging strategic industries and those with a more futuristic vision, thereby accelerating the formation of new quality productive forces. Later, at the annual Central Economic Work Conference, he further highlighted the importance of promoting industrial innovation through technological innovation, and to foster new industries, new business models and growth drivers, in particular, by leveraging ground-breaking and cutting-edge technologies, as part of our drive to develop new quality productive forces.

Kris Heylen, Katrien Depuydt, Carole Tiberius & Jesse de Does

BUILDING A SENSE INVENTORY FOR THE CENTRAL LEXICOGRAPHIC INFRASTRUCTURE OF DUTCH

Keywords Academic lexicography; lexicographic infrastructure; sense inventory; sense linking

Like other lexicographic and language institutes in Europe, the Dutch Language Institute (Instituut voor de Nederlandse Taal, INT) is progressively integrating the compilation and management of its individual lexicographic products into a single shared lexicographic infrastructure with the aim of increasing internal efficiency and enable new forms of use, research and development (Depuydt, Tiberius & Heylen 2026). Comparable infrastructural trajectories can be observed for Estonian (Koppel et al. 2019), Danish (Pedersen et al. 2018), or Slovene (Gantar 2020). For the INT, this integration is particularly challenging because its historical and contemporary dictionaries were originally largely developed as stand-alone projects and differ substantially in their internal structure.

Over the past two decades, the INT has systematically invested in addressing this challenge by developing a shared infrastructural backbone for Dutch lexicography. Since 2007, this has resulted in the creation of GiGaNT (Groot Geïntegreerd Lexicon van de Nederlandse Taal), a centralized lexicon providing unified lemmatisation principles, part-of-speech tagging, and form-related information for both historical and contemporary Dutch dictionaries and lexical databases (Ruitenberg, de Does, Depuydt 2010). GiGaNT enables the alignment of lexicographic resources at the lemma level across time periods and projects, and already supports the reuse of formal and morphological information in multiple lexicographic end products.

The present paper focuses on the next infrastructural step: the development of a sense inventory that extends this integration from the lemma level to the sense level so as to share and combine information that is currently distributed across databases. However, these resources, including the historical *Woordenboek der Nederlandsche Taal* (WNT), the contemporary general *Algemeen Nederlands Woordenboek* (ANW), and the phraseological *Woordcombinaties* differ not only in sense granularity, but also in their organising principles and descriptive focus. Therefore, a sense inventory is introduced as an intermediate semantic layer that allows these resources to be linked through core senses, to which definitions, corpus attestations, collocational patterns, usage information, and sense relations can be attached.

The objectives of this infrastructural development are twofold. Internally, the sense inventory is intended to support more efficient lexicographic workflows, enable systematic consistency checking across resources, and facilitate the reuse of semantic information in multiple dictionary products and services. Externally, it forms part of a publicly accessible lexicographic infrastructure that can be used for research on the Dutch lexicon, for the development of language learning materials, and for language-technology applications. In addition, the Dutch sense inventory is explicitly designed to contribute to and represent Dutch within the European lexicographic infrastructure ELEXIS, which will be substantially upgraded in the recently started Horizon Europe infrastructure project ELEXAI (2026-2029). Within ELEXAI, a collaborative and dynamic sense inventory will play a central role in the further development of the Dictionary Matrix and the European lexicographic knowledge graph.

The design of the Dutch sense inventory builds on earlier work at the INT, most notably the DiaMaNT project (Depuydt & de Does 2018), which established a semantic lexicon for

historical Dutch dictionaries by linking definitions across resources at a schematic level. This experience provides both methodological insights and reusable data for extending semantic integration to contemporary lexicography. At the same time, the current work takes into account international models and standards for lexicographic data modelling, including the DMLex data model (DMLex-1.0) and earlier proposals for unified sense-level organisation (e.g. Tavast et al. 2018).

The design of the Dutch sense inventory builds on earlier work at the INT, most notably the DiaMaNT project (Depuydt & de Does 2018), which established a semantic lexicon for historical Dutch dictionaries at a schematic level across resources by levering synonym definitions. This experience provides both methodological insights and reusable data for extending semantic integration to contemporary lexicography. At the same time, the current work takes into account international models and standards for lexicographic data modelling, including the DMLex data model (DMLex-1.0) and earlier proposals for unified sense-level organisation (e.g. Tavast et al. 2018).

This paper presents four interrelated components of the ongoing work. First, we introduce a preliminary data model for the Dutch sense inventory, which builds on the existing GiGaNT model while incorporating sense-level components already present in individual dictionaries. The model is designed to accommodate heterogeneous sense structures and to remain compatible with European interoperability efforts. Second, we discuss the principles for sense linking and sense lumping into core senses. These principles explicitly address large differences in sense granularity between dictionaries, the specific organising principles of *Woordcombinaties*, and the availability of explicit sense relations in the ANW, with attention to recurrent phenomena such as regular polysemy, for which earlier proposed (e.g. Pedersen et al. 2022) are evaluated and adapted.Third, we describe the compilation of a test dataset that serves both as a proof of concept and as an evaluation benchmark. The dataset includes lemmata with different parts of speech, varying degrees and types of polysemy, and contrasting treatments in the historical WNT, the contemporary ANW, and the pattern-based *Woordcombinaties*. It builds on previously curated material from DiaMaNT and on the Dutch component of the multilingual evaluation dataset for monolingual word sense alignment (Ahmadi et al. 2022). Finally, we discuss how this dataset will function as a baseline for future experiments in automatic sense linking, lumping and splitting using both open-weights models and commercial large language models.

References

Ahmadi, S., McCrae, J.P., Nimb, S. et al. (2020). A Multilingual Evaluation Dataset for Monolingual Word Sense Alignment. In N. Calzolari et al. (Eds.), *Proceedings of the Twelfth Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (pp. 3232–3242). European Language Resources Association (ELRA).

ANW: *Algemeen Nederlands Woordenboek*. Dutch Language Institute. Retrieved from http://hdl.handle.net/10032/tm-a2-k8

Depuydt, K., & De Does, J. (2018). The Diachronic Semantic Lexicon of Dutch as Linked Open Data. In N. Calzolari et al. (Eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA).

Depuydt, K., Tiberius, C., Heylen, K. (2026) Lexicography of the Dutch Language: From Dictionaries to a Lexicographical Infrastructure. *Reference Module in Social Sciences.* https://doi.org/10.1016/B978-0-323-95504-1.00427-0

DiaMaNT: *Diachroon seMantisch lexicon van de Nederlandse Taal.* Dutch Language Institute. Retrieved from http://hdl.handle.net/10032/tm-a2-q9

Filip, D., Jakubíček , M., Krek, S., McCrae, J. & Měchura, M. (Eds.). *Data Model for Lexicography Version 1.0*. OASIS Standard. Retrieved April 29, 2025, from https://docs.oasis-open.org/lexidma/dmlex/v1.0/dmlex-v1.0.html

Gantar, P. (2020). Dictionary of Modern Slovene: From Slovene Lexical Database to Digital Dictionary Database. *Rasprave: Časopis Instituta Za Hrvatski Jezik i Jezikoslovlje,* 46(2), 589–602. https://doi.org/10.31724/rihjj.46.2.7

GiGaNT-Molex (Version 2.0) [Data set]. Dutch Language Institute. Retrieved from http://hdl.handle.net/10032/tm-a2-u7

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. Ferreria, M. Jansen, I. Pereira, J. Kallas, M. Jakubíček, S. Krek & C. Tiberius (Eds.), *Electronic lexicography in the 21st century. Proceedings of the eLex 2019 conference* (pp. 434-452). Lexical Computing.

Pedersen, B. S., Nimb, S., Olsen, S. & Hartvig Sørensen, N. (2018). Combining Dictionaries, Wordnets and Other Lexical Resources– Advantages and Challenges. In I. Kernerman & S. Krek (Eds.), *Globalex Workshop Proceedings* (pp. 102-105). European Language Resources Association (ELRA).

Pedersen, B.S., Hau Sørensen N.C., Nimb, S., Flørke, 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 et al. (Eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (LREC) (pp. 51–60). European Language Resources Association (ELRA).

Ruitenberg, T., de Does, J. & Depuydt, K. (2016). Developing GiGaNT, a lexical infrastructure covering 16 centuries. In A. Dykstra & T. Schoonheim (Eds.), *Proceedings of the XIV Euralex International Congress* (pp. 468-476). Fryske Akademy.

Tavast, A., Koppel, K., Langemets, M. & Kallas, J. (2018). Unified Data Modelling for Presenting Lexical Data: The Case of EKILEX. In I. Čibej, V. Gorjanc, I. Kosem & S. Krek (Eds.), *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts* (pp.749-761). Ljubljana University Press.

WNT: *Woordenboek der Nederlandsche Taal*. Dutch Language Institute. Retrieved from https://hdl.handle.net/10032/tm-a2-g3

Woordcombinaties. Dutch Language Institute. Retrieved from https://hdl.handle.net/10032/tm-a2-t4

Contact information

Kris Heylen
Dutch Language Institute (Instituut voor de Nederlandse Taal – INT)
kris.heylen@ivdnt.org

Katrien Depuydt
Dutch Language Institute (Instituut voor de Nederlandse Taal – INT)
katrien.depuydt@ivdnt.org

Carole Tiberius
Dutch Language Institute (Instituut voor de Nederlandse Taal – INT)
carole.tiberius@ivdnt.org

Jesse de Does
Dutch Language Institute (Instituut voor de Nederlandse Taal – INT)
jesse.dedoes@ivdnt.org

Kris Heylen, Ilan Kernerman, Judit Freixa, Tarja Heinonen, Hae-Yun Jung, Henrik Lorentzen, Kilim Nam, Benjamin Rouxel, Ana Salgado, Miguel Sánchez Ibáñez, Petra Storjohann & Lars Trap-Jensen

THE ENEOLI MATRIX OF NEOLOGISM SELECTION CRITERIA

A Comparative Framework for Lexicographic Practice

Keywords Neology; lexicographic practice; dictionarization; selection criteria; comparative lexicography

Within the COST Action *European Network on Lexical Innovation* (ENEOLI[1]), Working Group 2, Task 2.5 on *Neology from a Lexicographic Perspective* has undertaken a collaborative effort to document, compare and systematize the criteria used by lexicographic projects when deciding on the inclusion of neologisms. Lexicographers and researchers presented their institutional policies and practices and integrated their inclusion and exclusion criteria in a shared analytical framework, resulting in a structured comparative matrix. This matrix is conceived as a point of reference: it enables professionals and researchers to compare methodologies and techniques across languages, institutional contexts and dictionary types, and it provides developers and data analysts with clearer insight into how tools and corpus analyses can support neology within the lexicographic workflow.

Our point of departure was the typology developed for Catalan (Freixa, 2022; Freixa Aymerich, 2022) that formalizes *dictionarizability* by structuring and weighing criteria such as frequency, dispersion, and linguistic integration. Building on this foundation, we extended the framework by integrating perspectives from other lexicographic traditions and languages, including Danish, Finnish, French, German, Korean, Portuguese, and Spanish. This cross-linguistic integration resulted in a comparative typology of selection criteria. The work of Task 2.5 was included in the design of the ENEOLI survey on methods and practices in lexicography and terminology (Kallas et al., 2025). The survey results provided additional insights from 27 distinct dictionary projects into how neologism selection criteria are implemented.

Taken together, our analysis identifies four interacting layers in neologism selection that form the basis of the comparative matrix:

First, evidence-based criteria can be distinguished by source: (1) usage data from corpora provide indicators such as minimal frequency, micro-diachronic developments in usage, and dispersion; (2) expert analysis evaluates word formation, semantic and pragmatic integration; (3) signals from the wider language community include user suggestions and dictionary logfiles; and (4) external lexical or terminological resources offer converging evidence. Although frequency is a dominant factor, it is rarely decisive in isolation but interpreted in interaction with these additional forms of evidence, that are especially relevant for languages and varieties with limited corpus coverage.

Second, workflow properties shape the operationalization of criteria, ranging from systematic candidate-list evaluation to case-by-case editorial judgments, and from inclusion- to exclusion-driven approaches.

[1] https://eneoli.eu/

Third, editorial governance structures influence authority and consistency, from top-down institutional guidelines to distributed decision making.

Fourth, a dictionary's overall functional orientation modulates the applicability of criteria. This includes macrostructural orientation (general monolingual vs. dedicated neologism dictionary), interest in documenting ephemeral formations, thematic focus, regional coverage, and positioning along a descriptive–proscriptive–prescriptive continuum.

By organizing these heterogeneous criteria along shared dimensions, and considering both inclusion and exclusion principles, the ENEOLI comparative matrix aims to enhance transparency in the documentation of lexical innovations and provide a reusable framework for comparing lexicographic practice, research and tool development. The matrix will be made publicly available via the ENEOLI repository.

References

Freixa, J. (2022). The Dictionarisation of Neologisms: The NADIC as a Model. In J. Freixa, M.I. Guardiola, J. Martines, M.A. Montané (Eds.), *Dictionarization of Catalan Neologisms* (pp. 15–37). Peter Lang Verlag.

Freixa Aymerich, J. (2022). Garbell: l'avaluador automàtic de neologismes catalans (Garbell: the automatic catalans neologisms analyzer). *Terminàlia*, 26, 7-16. https://publicacions.iec.cat/repository/pdf/00000323/00000022.pdf

Kallas, J., Koppel, K., Heylen, K., Kernerman, I., Ostroški Anic, A., Vezzani, F. & Arhar Holdt, Š. (2025). Neology in Practice: Lexicographic and Terminological Approaches to Lexical Innovation. In I. Kosem, M. Jakubícek, M. Medved, K. Zgaga, Š. Arhar Holdt, T. Munda & A. Salgado (Eds.), *Proceedings of the eLex 2025 conference* (pp. 652–672). Lexical Computing CZ.

Nam, K., Lee, S., Jung, H.-Y & Choi, J. (2016). The life and death of neologisms. On what basis shall we include neologisms in the dictionary? In T. Margalitadze & G. Meladze (Eds.), *Proceedings of the XVII EURALEX International Congress: Lexicography and Linguistic Diversity* (pp. 389–393). Ivane Javakhishvili Tbilisi State University.

Nam, K., Soojin, L. & Jung, H.-Y. (2020). The Korean Neologism Investigation Project: Current Status and Key Issue. *Dictionaries: Journal of the Dictionary Society of North America,* 41(1), 105-29. https://doi.org/10.1353/dic.2020.0007

Sánchez Ibáñez, M. & Maroto, M. (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), 358-81. https://doi.org/10.1093/ijl/ecab015

Storjohann, P. (2024). IDS-Neo 2020+: A Novel Resource for New German Words in Use. *International Journal of Lexicography,* 37(4), 389-403. https://doi.org/10.1093/ijl/ecae015

Trap-Jensen, L. (2020). Language-Internal Neologisms and Anglicisms: Dealing with New Words and Expressions in The Danish Dictionary. *Dictionaries: Journal of the Dictionary Society of North America,* 41(1), 11-25. https://doi.org/10.1353/dic.2020.0002

Dictionaries:

Den Danske Ordbog. Det Danske Sprog- og Litteraturselskab. Retrieved February, 25, 2026, from https://ordnet.dk/ddo

Dicionário da Língua Portuguesa. Academia das Ciências de Lisboa. Retrieved February, 25, 2026, from https://dicionario.acad-ciencias.pt/

Kielitoimiston sanakirja. Kotimaisten kielten keskus. Retrieved February, 25, 2026, from https://www.kielitoimistonsanakirja.fi/#/

Le Petit Robert de la langue française. Dictionnaires Le Robert-SEJER. Retrieved February, 25, 2026, from https://www.lerobert.com/

Contact information

Kris Heylen
Dutch Language Institute (Instituut voor de Nederlandse Taal, INT)
kris.heylen@ivdnt.org

Ilan Kernerman
Lexicala by K Dictionaries
ilan@lexicala.com

Judit Freixa
Universitat Pompeu Fabra
judit.freixa@upf.edu

Tarja Heinonen
Institute for the Languages of Finland (Kotimaistenkielten keskus)
tarja.heinonen@kotus.fi

Hae-Yun Jung
Kyungpook National University
haeyun.jung.22@gmail.com

Henrik Lorentzen
Society for Danish Language and Literature
HL@dsl.dk

Kilim Nam
Yonsei University
nki@yonsei.ac.kr

Benjamin Rouxel
Éditions Le Robert
brouxel@lerobert.com

Ana Salgado
Academia das Ciências de Lisboa
anacastrosalgado@gmail.com

Miguel Sánchez Ibáñez
University of Valladolid
miguelsanchez@uva.es

Petra Storjohann
Leibniz-Institut für Deutsche Sprache, Mannheim
storjohann@ids-mannheim.de

Lars Trap-Jensen
Society for Danish Language and Literature
ltj@dsl.dk

Acknowledgements

This paper is based on work from COST Action CA22126 – European Network on Lexical Innovation (ENEOLI - eneoli.eu), supported by COST (European Cooperation in Science and Technology).

Louise Holmer

NEOLOGY MILLS?

AI-Influenced Newspaper Articles and their Implications for Lexicology and Lexicography

Keywords GenAI-translations; lexicography; neologisms; SAOL 15 2026; SO 2026; Swedish

Over the past several years, the use of generative AI (GenAI) has increased significantly in news media. Newspaper articles can nowadays be produced, revised or translated with the assistance of GenAI. In some cases, this is explicitly stated in the text, thereby increasing transparency for the reader.

The aim of this project is to present a study on newspaper articles that have been translated using GenAI from English into Swedish and subsequently revised by the newspaper's editors. Both the English source texts and the Swedish versions are available online. This makes it possible to conduct manual analyses as well as computational analyses regarding lexical units, phraseology and syntactic features, as well as the overall style and register of the Swedish target texts.

Since modern dictionaries intended for human users rely heavily on text corpora, which as a rule include newspaper articles, the presence of GenAI-texts in contemporary corpora and the varying qualities of such texts raise many questions for lexicographers. If AI-generated texts are of sufficiently high quality, containing conventionalized vocabulary as well as neologisms and nonce-words, mostly well-formed phrases and collocations, and displaying overall coherence, their inclusion may not pose a major problem for lexicographers. Whether this is the case is of course an empirical question, which I address in this study.

This study has a lexicographical approach. It involves the two modern Swedish dictionaries published by The Swedish Academy, namely SAOL 15, 2026 (*The Swedish Academy Glossary*) and SO 2026 (*The Contemporary Dictionary of the Swedish Academy*) (Holmer & Blensenius, 2025; Sköldberg et al, 2025). These two dictionaries are compiled at the University of Gothenburg with funding from the Swedish Academy. Both rely upon corpus linguistic methods and language technology (see e.g., Borin et al. 2025), and the text base is dominated by newspaper articles.

My working hypothesis for this study is that the GenAI-translated Swedish texts may contain unexpected lexical items (including possible nonce-formations and neologisms), unexpected phrasing, and non-standard syntactic structures in comparison with original Swedish non-translated texts. To test this hypothesis, I compiled a corpus of newspaper articles published in the daily newspaper *Dagens Nyheter* (DN) during late 2025 and early 2026. The articles were originally published in *The New York Times* and were explicitly identified as having been translated from English into Swedish using AI and subsequently reviewed by editors at DN.

The collected texts are first examined manually and subsequently prepared for computational lexical analysis using tools provided by Språkbanken (2026), a research infrastructure for language data. The results are then compared with previous studies of word frequency and lexical profiling in Swedish, including contrastive studies on *translationese* (Gellerstam, 1986; cf. Volansky et al., 2015). Finally, potential implications for lexicographically oriented scholars are identified and discussed.

One could ask whether it matters if a language acquires nonce-words and neologisms through AI-generated texts, and if so, to whom it matters. For lexicologically and lexicographically oriented scholars, it probably does (see e.g., Poix & Shevchenko, 2025; a post on Language Log from May 2026 discusses the status of "actorive"). Lexicographers must carefully, and often manually, examine lexical items identified as potential neologisms to determine whether such items should be included in a dictionary, and if so, in the position of for example headwords or quotations.

References

Borin, L., Forsberg, M., Hammarstedt, M., Holmer, L. & Matsson, A. (2025). Korp: Språkbanken's word research platform. In D. Dannélls, K. Blensenius & L. Borin (Eds.), *Sixty years of Swedish computational lexicography* (pp. 175–194). De Gruyter. https://doi.org/10.1515/9783111577234-010

Gellerstam, M. (1986). Translationese in Swedish novels translated from English. In L. Wollin & H. Lindqvist (Eds.), *Translation Studies in Scandinavia* (pp. 88–95). CWK Gleerup.

Holmer, L. & Blensenius, K. (2025). SAOL: a Swedish dictionary for all times. In D. Dannélls, K. Blensenius & L. Borin (Eds.), *Sixty years of Swedish computational lexicography* (pp. 27–52). De Gruyter. https://doi.org/10.1515/9783111577234-003

Poix, C. & Shevchenko, N. (2025). The challenge of AI-generated neology. In I. Kosem, M. Jakubíček, M. Medveď, K. Zgaga, Š. Arhar Holdt, T. Munda & A. Salgado (Eds.), *Proceedings of the eLex 2025 conference* (pp. 322–335). Lexical Computing CZ.

Rips, M. (2026). "Actorive"?? Language Log 2026-05-20, posted by Mark Liberman. Retrieved May 25, 2026, from https://languagelog.ldc.upenn.edu/nll/?p=73565

SAOL 15 = *Svenska Akademiens ordlista över svenska språket/The Swedish Academy Glossary*, 15th edition (2026). Norstedts.

Sköldberg, E., Blensenius, K. & Holmer, L. (2025). SO: The Swedish contemporary dictionary. In D. Dannélls, K. Blensenius & L. Borin (Eds.), *Sixty years of Swedish computational lexicography* (pp. 53–82). De Gruyter. https://doi.org/10.1515/9783111577234-004

SO 2026 = *Svensk ordbok utgiven av Svenska Akademien/The Contemporary Dictionary of the Swedish Academy* (2026). Retrieved May 25, 2026, from https://svenska.se/?activeTab=so

Språkbanken 2026. Retrieved May 25, 2026, from https://spraakbanken.gu.se/en

Volansky, V., Ordan, N. & Wintner, S. (2015). On the features of translationese. *Digital Scholarship in the Humanities*, 30(1), 98-118. https://doi.org/10.1093/llc/fqt031

Contact information

Louise Holmer
Språkbanken, Department of Swedish, Multilingualism, Language Technology University of Gothenburg
louise.holmer@svenska.gu.se

Željko Jozić, Marijana Horvat & Martina Horvat

DIGITIZATION OF ADAM PATAČIĆ'S MANUSCRIPT DICTIONARY

Keywords Adam Patačić; Dictionarium latino-illyricum et germanicum; dictionary; digitization; Kajkavian Croatian literary language

Adam Patačić is a Croatian lexicographer (1716–1784), Archbishop of Kalocsa and President of the Royal Council of the University of Buda and a promoter of music with his orchestra and choir. His extensive Dictionarium latino-illyricum et germanicum remained in manuscript, written after the example of Nomenclator omnium rerum by Hadrian Junius (Samardžija, 2019, p. 63).

The mentioned dictionary is a completed manuscript work written in ink, the pages are numbered, and on the inside title page the author is signed as D. ADAMUS L. B. PATACHICH De Zajezda. Although the manuscript does not contain information on the title page indicating the year the work dates from, the data relevant to the dating of the dictionary can be read from the preface, as well as from the title page. According to the data on the title page, the author's name is accompanied by a list of his titles and honors, some of which he achieved only towards the end of his life. In support of this dating, and according to the numbered third page of the dictionary, it can be concluded that the work was bound and completed in the late seventies or early eighties of the 18th century. According to the data from the preface, it can be stated with certainty that the dictionary was created in the period from 1772 to 1779 (Jonke, 1949, pp. 95–96). Patačić's dictionary was written in an era that, together with the previous century, is considered the golden age in the history of the Kajkavian Croatian literary language. It was at this time that major Kajkavian lexicographical works were created – Belostenec's, Jambrešić's and Patačić's dictionaries, as well as the first orthography and the first manuscript Kajkavian grammar (Štebih Golub, 2013, p. 258).

The dictionary is conceptually structured with a macrostructure of 13 thematic areas and several sub-areas, with encyclopedic interpretations in Latin. The headword side is Latin, and the explanations are given in Croatian, i.e. in the Kajkavian literary language, while the equivalent is also given in German. The dictionary contains a rich treasury of Kajkavian literary lexis (for example, kinship, zoological and botanical terminology), as well as the Kajkavian literary language of the 18th century.

The preface contains XIII pages, while the content consists of V pages. The dictionary material covers 1034 numbered pages. The dictionary is divided into four major parts. The first part entitled De Deo Spiritibus, Coelo, Elementi set Homine comprises 13 chapters on 220 pages and provides general information about the extralinguistic reality. The second part is entitled De iis, quae ad Hominis vitam tum recte agendam, tum tuendam pertinent: sive Sacrorum Praesides, Civilia et Militaria, and contains 15 chapters (pp. 221–415) and covers political and military duties, literature, music, and art. The third part is entitled Oeconomica and contains 14 chapters (pp. 416–734) and relates to the economy. The fourth part Technica (pp. 735–1034) consists of 16 chapters and relates to technical sciences.

Although Patačić's dictionary was not used as a corpus in the preparation of the Academy's historical multi-volume Dictionary of the Croatian or Serbian Language (1880–1976), it served as a corpus in the preparation of the Dictionary of the Croatian Kajkavian Literary Language (1984–present, kajkavski.hr) where, in structuring the dictionary entry, confirmations from

well-known Kajkavian dictionaries of the 17th and 18th centuries are cited, along with Patačić's manuscript dictionary (Brlobaš and Horvat, 2019, p. 146).

The largest study of the dictionary is Ljudevit Jonke's Dikcionar Adama Patačića (1949) and the more recent one is Marijana Horvat's research on musical terminology in Croatian pre-revival dictionaries (Horvat, 2018, pp. 437–454).

Due to its exceptional importance for the Croatian language in general and the Kajkavian literary language, at the beginning of 2026, at the initiative of the director of the Institute for the Croatian Language, Željko Jozić, and with the help of collaborators on the Dictionary of the Croatian Kajkavian Literary Language project, and with other colleagues from the Department of the History of the Croatian Language, a project was launched to study Patačić's manuscript dictionary. The transliteration of the work has begun with the aim of further studying the extensive and the last historically extremely important Kajkavian dictionary. In the presentation, we will talk about the principles of digitization and the problems we encounter.

We are digitising Adam Patačić's handwritten Dikcionar, a manuscript exceeding 1,000 pages, by combining philological transcription and handwritten text recognition (HTR) in Transkribus. As initial Ground Truth, we reused the manuscript text transcribed by Ljudevit Jonke in his study 'Dikcionar' Adama Patačića: studija iz hrvatske kajkavske leksikografije. This transcription corresponds to 25 manuscript pages of the Dikcionar. The text was imported into Transkribus and manually aligned with the corresponding manuscript images and baselines, yielding a first controlled Ground Truth set. Our very first training attempt—based on an incomplete subset of this Jonke-derived material and without an appropriate base model—produced poor recognition performance (Transkribus-reported CER 58.61%), showing that substantial script-specific adaptation was still needed. We therefore introduced transfer learning and retrained the model using the publicly available base model "Cosimo Bartoli's Italian Humanistic and Cursive Scripts (1562–1572)" together with the Jonke-based Ground Truth. This markedly improved recognition, lowering the Transkribus CER to 11.52%. Since our research use case tolerates several predictable early-modern graphemic ambiguities, we additionally evaluated performance under a normalisation that ignores case and treats u/v, s/ſ, and æ/ae as equivalent. Under these conditions, an independent CER calculation yielded 6.01%, which we consider an excellent and practically usable starting point for large-scale digitisation. The next phase will proceed iteratively: a team of researchers will expand Ground Truth by correcting model output on newly processed pages, while the model will be periodically retrained to further improve robustness across the manuscript and to support subsequent linguistic and lexicographic analysis of the Dikcionar. This digitisation also marks an exceptional pilot step within the MONOGRAF project (Development and Implementation of a Model for Normalising the Orthography of Old Texts Printed in Latin Script), which otherwise focuses on printed texts but has, on a trial basis, extended its workflow to a handwritten manuscript in this case.

References

Brlobaš, Ž., Horvat, M. (2019). Metodologija leksikografske obradbe Rječnika hrvatskoga kajkavskoga književnog jezika. *Jezikoslovni zapiski*, 25(1). 143–164.

Horvat, M. (2018). Glazbeno nazivlje u hrvatskim dopreporodnim rječnicima. *Rasprave Instituta za hrvatski jezik*, 44(2), 437–454.

Jonke, Lj. (1949). Dikcionar Adama Patačića. *Rad JAZU*, 275(2), 71–175.

Samardžija, M. (2019). *Hrvatska leksikografija od početaka do kraja XX. stoljeća*. Matica hrvatska.

Štebih Golub, B. (2013). Kajkavski hrvatski književni jezik u 17. i 18. stoljeću. In Katičić, R., Lisac, J. (Eds.) *Povijest hrvatskoga jezika 3* (pp. 221–263). Croatica.

Contact information

Željko Jozić
Institut za hrvatski jezik, Zagreb
zjozic@ihjj.hr

Marijana Horvat
Institut za hrvatski jezik, Zagreb
mhorvat@ihjj.hr

Martina Horvat
Institut za hrvatski jezik, Zagreb
mshorvat@ihjj.hr

Acknowledgements

This article was produced within the framework of the project "Development and Implementation of a Model for Normalising the Orthography of Old Texts Printed in Latin Script (MONOGRAF)", 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.

Hae-Yun Jung, Jinsan An, Kilim Nam & Jun Lee

FROM NEOLOGISM COLLECTION TO LEXICOGRAPHIC DESCRIPTION

A Corpus-LLM-Lexicographer Triangulation Framework

Keywords Korean neologisms; methodological triangulation; corpus; large language models (LLMs); Retrieval-Augmented Generation (RAG)

This paper introduces the *Korean Neologism Observatory* (KNO), which has been developed through the methodological triangulation (Baker & Egbert, 2016) of "corpus," "large language model (LLM)," and "lexicographic experts." In line with recent attempts to introduce generative AI into the collection and description of neologisms (e.g., Nam et al., 2025; Tomaszewska et al., 2025; Widmann, 2025), this study discusses the innovation brought about by an LLM-based triangulation methodology and the technical limitations of LLM-based dictionary compilation.

Although methodologies for investigating Korean neologisms have developed from manual to semi-automatic extraction, they remain labour-intensive, as is the case for other languages such as English, German, and Spanish. KNO is a human-in-the-loop system developed to enhance the productivity of neologism investigation by adding an LLM-assisted neologism identification and description stage to the existing semi-automatic extraction methodology. The main outputs are made publicly available via GitHub (https://github.com/KorNEO/KNO).

Through the KNO pipeline presented in Figure 1, the annual collection of Korean neologisms conducted since 1994 has been subdivided into monthly neologism collections, and the scope of neologism collection has also been expanded to encompass user discourse such as YouTube comments. From the perspective of Stefanowitsch & Flach (2017), the News Article Corpus represents linguistic input (corpus-as-input), while the YouTube Comment Corpus represents both linguistic input and linguistic output (corpus-as-input/output) that provides data for the observation of speakers' direct creation and use of neologisms. Unlike previous neologism research relying on newspaper corpora, encompassing user discourse has made it possible to analyse the initial topic and genre of occurrence, diffusion paths and intensity, and the number of speakers using a particular neologism. The LLM-assisted neologism identification and description stage uses the Retrieval-Augmented Generation (RAG) technique, which utilizes corpus examples and the existing Neologism DB as reference materials to minimize the limitations of LLMs, including non-determinism, non-reproducibility, and hallucinations, while complementarily identifying neologisms that humans may have omitted when reviewing candidate lists. The usage patterns of collected neologisms are then monitored monthly, their diffusion types classified according to durability and scope of use, and their candidacy for dictionary inclusion evaluated using quantitative and qualitative criteria.

In the experimental process of collecting and describing approximately 740 morphological neologisms (January 1 to December 31, 2025), additional technical limitations of LLM-based lexicography were observed. LLMs showed not only low performance in resolving lexical semantic ambiguity, especially in differentiating the meanings of homonyms and polysemous words, but also limitations in capturing the socio-cultural meanings of neologisms reflecting rapidly changing political discourse. When comparing the results of LLM-assisted neologism identification (GPT-5.2, with RAG) against final judgments by human experts through a confusion matrix, the precision rate was approximately 18% and the recall rate approximately 96%,

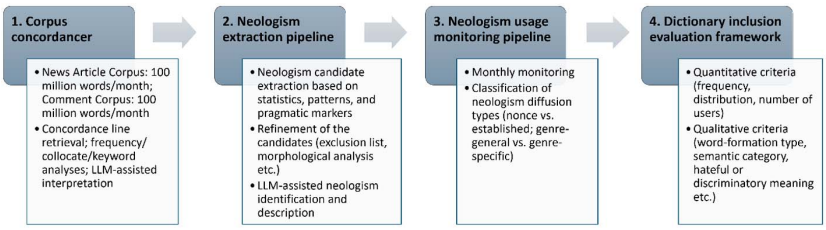


FIGURE 1 Structure of the KNO pipeline

indicating the effectiveness of LLM at capturing most actual neologisms from a large candidate pool. Nevertheless, there were also cases where human experts identified items as neologisms that the LLM had classified as non-neologisms (Example 1), while in other instances, the LLM incorrectly classified free combinations or proper nouns as neologisms.

> (1) 두쫀쿠 *twucconkhwu* < 두바이 쫀득 쿠키 *twupai ccontuk khwukhi* 'Dubai chocolate cookie'; 야차 룰 *yacha lwul* 'yaksha rule'; 연프 재질 *yenphu caecil* < 연애 프로그램 재질 *yenay phulokulaym caecil* 'dating show vibes'

Furthermore, in the LLM-assisted neologism description task, the RAG-based generation contributed to refining the metalanguage of lexicographic definitions (e.g., "a figurative term referring to X"). However, for neologisms whose word formation is inherently motivated by its figurative sense, such as 경제 단두대법 *kyengcey tantwutaypep* 'economic guillotine bill', the LLM merely summarized evaluative statements found in corpus examples and failed to reflect the metaphorical meaning of 단두대 *tantwutay* 'guillotine' in the definition.

Building on these findings, this study first examines the similarities and differences between human language and LLM language; it then discusses the significance of human linguistic intuition in determining neologicity (Lombard et al., 2021, 2024) and the need to reinterpret the artistry and craftsmanship (Landau, 1984) of the lexicographer in the dictionary description of low-frequency words within the context of LLM-assisted neologism lexicography. To ensure the replicability and falsifiability (McEnery & Brezina, 2022) of neologism research, this study proposes a system that automatically logs the configuration, input, output, and timestamp of each pipeline module, thereby enabling the tracking of both the collection path of individual neologisms and the path by which they have been missed.

References

Baker, P., & Egbert, J. (Eds.) (2016). *Triangulating methodological approaches in corpus linguistic research*. Routledge.

Landau, S. I. (1984). *Dictionaries: The art and craft of lexicography*. Cambridge University Press.

McEnery, T., & Brezina, V. (2022). *Fundamental principles of corpus linguistics*. Cambridge University Press.

Nam, K., An, J., & Lee, S. (2025). A Triangulational Approach to Korean Neologism Detection: Corpora, Large Language Models, and Human Expertise in Neologisms 2023. *Korean Linguistics*, 108, 205-238. https://doi.org/10.20405/kl.2025.08.108.205

Lombard, A., Huyghe, R., & Gygax, P. (2021). Neological intuition in French: A study of formal novelty and lexical regularity as predictors. *Lingua*, 254, 103055. https://doi.org/10.1016/j.lingua.2021.103055

Lombard, A., Huyghe, R., & Gygax, P. (2024). Morphological productivity and neological intuition. *Glossa Psycholinguistics*, 3(1). https://doi.org/10.5070/G60111291

Stefanowitsch, A., & Flach, S. (2017). The corpus-based perspective on entrenchment. In H.-J. Schmid (Ed.), *Entrenchment and the psychology of language learning: How we reorganize and adapt linguistic knowledge* (pp. 101-127). De Gruyter Mouton. https://doi.org/10.1515/9783110341423-006

Tomaszewska, A., Czerski, D., Żuk, B., & Ogrodniczuk, M. (2025). NeoN: A Tool for Automated Detection, Linguistic and LLM-Driven Analysis of Neologisms in Polish. In M. H. Lees, W. Cai, S. A. Cheong, Y. Su, D. Abramson, J. J. Dongarra & P. M. A. Sloot (Eds.), *International Conference on Computational Science* (pp. 318-326). Cham: Springer Nature Switzerland.

Widmann, T. (2025). A pipeline for automated dictionary creation with optional human intervention. In I. Kosem, M. Jakubíček, M. Medveď, K. Zgaga, Š. Arhar Holdt, T. Munda & A. Salgado (Eds.), *Electronic lexicography in the 21st century (eLex 2025): Intelligent lexicography. Proceedings of the eLex 2025 conference* (pp. 33-43). Lexical Computing CZ s.r.o.

Contact information

Hae-Yun Jung
Kyungpook National University
haeyun.jung.22@gmail.com

Jinsan An
Kyungpook National University
san@knu.ac.kr

Kilim Nam
Yonsei University
nki@yonsei.ac.kr

Jun Lee
Yonsei University
leejun0624@gmail.com

Ilan Kernerman, Federica Vezzani, Anna Vacalopoulou, John Humbley, Ana Salgado, David Lindemann, Michael Rosner, Raquel Silva & Onorina Botezat

DESIGNING THE MACROSTRUCTURE AND MICROSTRUCTURE OF THE ENEOLI MULTILINGUAL GLOSSARY OF NEOLOGY

Keywords Neology; multilingual glossary; macrostructure; microstructure; terminology

The first objective of the COST Action ENEOLI – European Network on Lexical Innovation[1] – is to "[d]efine the core terminology of neology conceived as a discipline through the creation of a born-digital specialised multilingual glossary."[2] This resource, which will be freely accessible to professionals, researchers, students and the public, is being developed collaboratively by the seven task teams composing ENEOLI's Working Group 1 (WG1). Within this framework, Task 1.4 (T1.4) oversees designing the glossary's macrostructure (the organisation of entries) and microstructure (the data categories used within entries), sourced from ENEOLI Wikibase (Lindemann & Salgado, 2025) and edited as a structured dataset. This paper highlights the main aspects of the research, development and implementation involved in this task, and showcases the final architecture and sample concepts and terms.

To ensure optimal multilinguality, precision, comprehensiveness, interoperability and usefulness, the glossary's macrostructure needs to accommodate equal inclusion of any language, and the microstructure components need to cater to all languages. The idea of a simpler semi-multilingual model was rejected to avoid prioritizing a single pivot language representing the full entry content that would leave other language equivalents available only as lemmata. While English is accorded a lingua franca status that connects all the languages and functions as an initial reference to guide the term selection, French constitutes the language of the starter corpus from which the initial terms were retrieved, and other languages may contribute more concepts that are particularly salient within their own linguistic and cultural traditions.

Another key challenge was to explore the similarities and discrepancies between terminological and lexicographic approaches and features, and to harmonize them coherently and efficiently. This challenge stems from the different functions of each, as lexicography deals with lexical units, whereas terminology aims at identifying and possibly structuring concepts. Thus, while this pioneering glossary is primarily a terminological resource, it is also regarded as part of the lexicographic domain, especially through its use of words that account for lexical innovation.

As part of this research, T1.4 members experimented with compiling glossary entries in ten additional languages[3] and investigated which elements to include and how to arrange and

1 https://eneoli.eu/

2 https://cost.eu/actions/CA22126/

3 Albanian (Anila Çepani), Arabic (Amal Haddad Haddad; Lina Sager Feghali), Greek (Anna Vacalopoulou), Icelandic (Ágústa Þorbergsdóttir and Kristin Ingibjörg Hlynsdóttir), Korean (Kilim Nam), Lithuanian (Sigita Rackevičienė and Giedrė Valūnaitė-Oleškevičienė), Portuguese (Ana Salgado), Romanian (Corina Veleanu), Turkish (Neslihan Önder Özdemir), Ukrainian (Rusudan Makhachashvili), in addition to English and French.

TABLE 1 Entry layers and data categories and sub-categories; * indicates mandatory category

layer	data category	data sub-category
Concept	Concept ID*	
	English Gloss*	
		Gloss
	Subject Field*	
		Subject Subfield
	Associated Concept ID	
	Note	
Language	Language*	
		Language in Working Language*
		Language ID*
	Definition*	
	Note	
Term	Term*	
		Term ID*
	IPA	
		Alternative Phonetic Transcription
	Audio Pronunciation	
	Variant Spelling	
	Variant Script	
	Part of Speech*	
		Grammatical Gender
		Grammatical Number
	Attestation	
		Source
		Hyperlink
		Dating
	Note	
	Associated Term	
		Associated Term ID

interlink them. In accordance with ISO 16642:2025,[4] the macrostructure was ultimately designed as a three-tiered network articulated into Concept, Language, and Term layers, and the microstructure went through further rounds of drafting-discussion-redrafting, to evaluate and compare the potential features and define the final structure and components. Overall, approximately seventy potential data categories were sketched in and reviewed, and several provisional lists were drafted and revised, ending with the microstructure including 32 data categories and sub-categories, of which ten are mandatory and the rest are optional. Table 1

[4] ISO 16642:2025 Management of terminology resources – Terminological markup framework. https://iso.org/standard/87351.html

presents the three-layer entry structure and the data categories. Full integration of the final macrostructure and microstructure into the working platform was concluded in mid-2026, enabling subsequent compilation in all languages.

References

Lindemann, D. & Salgado, A. (2025). ENEOLI Wikibase: A Collaborative Working Platform for the European Network on Lexical Innovation. In I. Kosem, M. Jakubiček, M. Medved', K. Zgaga, Š. Ahar Holdt, T. Munda & A. Salgado (Eds.), *Electronic lexicography in the 21st century (eLex 2025): Intelligent lexicography. Proceedings of the eLex 2025 conference* (pp. 289–300). Lexical Computing CZ s.r.o. https://elex.link/elex2025/wp-content/uploads/eLex2025-19-Lindemann Salgado.pdf

Contact information

Ilan Kernerman
Lexicala by K Dictionaries
ilan@lexicala.com

Federica Vezzani
Università degli Studi di Padova
federica.vezzani@unipd.it

Anna Vacalopoulou
Athena Research Center
avacalop@athenarc.gr

John Humbley
Université Paris Cité
humbley.john@orange.fr

Ana Salgado
Universidade do Porto
anacastrosalgado@gmail.com

David Lindemann
University of the Basque Country
david.lindemann@ehu.eus

Michael Rosner
University of Malta
mike.rosner@um.edu.mt

Raquel Silva
NOVA University
raq.asilva@gmail.com

Onorina Botezat
Bucharest University of Economic Studies
onorina.botezat@rei.ase.ro

Acknowledgements

ENEOLI is a COST Action network funded by the EU's Horizon 2020 Framework Programme. COST (Cooperation in Science and Technology) is a funding agency for research

and innovation networks, connecting research initiatives across Europe and enabling scientists to grow their ideas by sharing them with their peers, thus boosting their research, career and innovation.

Simon Krek, Iztok Kosem & Polona Gantar

BRIDGING LEXICOGRAPHY AND NLP

The Architecture and Application of the Digital Dictionary Database of Slovene (DDDS)

Keywords Relational database; Slovene language; data modelling; responsive dictionaries; NLP resources

In the era of data-driven linguistics, the transition from traditional, human-oriented lexicography to machine-readable and interoperable language resources is paramount. The Digital Dictionary Database of Slovene (DDDS), developed by the Centre for Language Resources and Technologies (CJVT) at the University of Ljubljana, represents a paradigm shift in how national lexicographical data is curated, stored, and disseminated. Rather than treating dictionaries as isolated digital products, the DDDS functions as a centralized, relational infrastructure designed to support a multi-layered ecosystem of language resources—ranging from the Thesaurus of Modern Slovene to the Collocations Dictionary and morphological database Sloleks. This paper will describe the structural design, the underlying data model, and the multifaceted API-based application of the DDDS, highlighting its role in the "responsive dictionary" framework.

At its core, the DDDS is built on a PostgreSQL database managed via the Django Object-Relational Mapping (ORM) framework. Unlike traditional lexicographical databases that often rely on XML hierarchies, the DDDS employs a highly normalized relational model that prioritizes the atomization of linguistic data. The model is organized into several clusters, with the primary entities being Lexical Units and Senses. A key innovation in the DDDS is its approach to "Lexemes," which correspond to tokens in a corpus but are treated as distinct structural components within the database. This allows for the precise representation of complex multi-word expressions, compounds, and syntactically varied forms. By utilizing a single primary key system and Django's "content types" and "generic relations," the database maintains a flexible architecture where additional metadata—such as dictionary-specific labels or administrative information—can be associated with any table without necessitating disruptive schema changes. This modularity ensures that the database remains a "living" resource capable of accommodating various lexicographical workflows.

The DDDS serves as the "backbone" for a suite of modern Slovene resources. Most notably, it powers the Thesaurus of Modern Slovene, which introduced the concept of the responsive dictionary. In this model, the dictionary is not a static text but an evolving dataset where automatically generated candidates are continuously refined through expert validation and crowdsourcing. The database manages these dynamic relationships, tracking synonymy, sense disambiguation, and collocation frequency extracted from the Gigafida reference corpus.

A critical component of the DDDS project is its commitment to the FAIR (Findable, Accessible, Interoperable, and Reusable) principles. This is achieved through a comprehensive REST API that allows external developers, researchers, and NLP systems to interact with the database in real-time. The API design distinguishes between read-only public routes and restricted read-write routes.

A significant modern application of the DDDS is its role in the PoVeJMo project, which aims to optimize Large Language Models for the Slovene language. By providing high-quality, structured grammatical and lexical data, the DDDS enables models to learn the nuances of Slovene

grammar and semantics more effectively than they would from unstructured web-crawl data alone. The extraction of sense-specific examples and sense-relation mappings from the database provides a "gold standard" dataset for training models in tasks such as Natural Language Inference (NLI) and Word Sense Disambiguation (WSD).

The development of the DDDS highlights a shift toward a "digital-first" lexicographic workflow. Traditional dictionary writing involved drafting entries that were later converted to digital formats. In the DDDS framework, the database is the dictionary. The interface used by lexicographers—a new dedicated online editor—is integrated with the database, ensuring that updates are immediately reflected in the underlying data structure. The Digital Dictionary Database of Slovene demonstrates that a central, well-structured database is essential for the survival and growth of a national language in the digital age. By decoupling the linguistic data from the final presentation layer (the web interface), CJVT aims to create a robust infrastructure that supports both human users and automated systems. Future developments involve the expansion of the data model to include more granular semantic relations and the continued integration of crowdsourced feedback, ensuring that the DDDS remains the authoritative, yet dynamic, reference for the modern Slovene language.

In the final paper we will describe the specialized data model with REST API, the online editor, and various online visualizations of DDDS data.

Contact information

Simon Krek
Jožef Stefan Institute
simon.krek@guest.arnes.si

Iztok Kosem
University of Ljubljana
iztok.kosem@cjvt.si

Polona Gantar
University of Ljubljana
apolonija.gantar@guest.arnes.si

Amila Kugic

ANALYSIS OF SYNTHETIC CLINICAL NEOLOGISMS

A Corpus-Based Analysis of Large Language Model-Generated Content in the Context of Artificial Intelligence-Assisted Clinical Documentation

Keywords Natural language processing; artificial intelligence; terminology; neologism; controlled vocabulary

With generative artificial intelligence (AI) tools becoming more widely available and integrated into clinical environments, physicians are more frequently exploring ways to integrate AI into their clinical documentation and decision-making processes (Brender et al., 2025). Research already shows that large language models (LLMs) have the capability for summarization and analysis of clinical documents but carry also risks towards introduced hallucinated or biased content that can potentially skew further diagnostic steps and treatment options (Busch et al., 2025). Previous research suggested that the addition of neologisms to texts ended up decreasing machine translation quality by an average of 43% (Zheng et al., 2024).

This current investigation examined synthetically LLM-generated German clinical corpora and the use of neologisms and lexical patterns that differ from established controlled vocabularies, such as the Unified Medical Language System (UMLS) and SNOMED CT. Four synthetic German clinical training datasets were generated using GPT-4o and GPT-5 models. Each dataset comprising 900 snippets was constructed to compare zero-shot and few-shot prompting strategies, where both prompts instructed the generation of short, approximately 100-character length clinical-style snippets. LLM outputs were designed to reflect realistic clinical documentation features, such as abbreviated syntax, incomplete sentences, and laboratory-style reporting. Words or multiword expressions absent from reference biomedical corpora were flagged as candidate neologisms. These were then analyzed based on their morphological and structural properties.

To assess semantic relatedness between candidate neologisms and established biomedical terminology, embedding-based similarity analysis was performed using a pretrained multilingual transformer model (sentence-transformers/paraphrase-multilingual-mpnet-base-v2). Each candidate term and each controlled vocabulary entry was converted into a vector representation. The nearest-neighbor approach via cosine similarity allowed each candidate form to be linked to its most semantically similar standardized form. The method compared candidate terms that closely match existing medical concepts linked in terminologies, and those with low similarity scores, which were more likely to represent hallucinated expressions. The similarity scores were furthermore combined with frequency information to highlight commonly occurring but semantically distant forms for further analysis. This embedding-based comparison supported also frequency-based filtering of candidate terms for manual validation and review.

Preliminary analysis of synthetically generated German clinical narrative snippets revealed that the majority of candidate neologisms were not entirely novel lexical terms with reference to SNOMED CT and UMLS, but rather orthographic variants, abbreviation modifications, and fragmented forms derived from existing biomedical terminology. Co-occurrence analysis demonstrated clustering around conventional documentation templates when few-shot prompting

was implemented in the synthetic generation pipeline. This suggested that neologisms primarily emerged through recombination and modification of established documentation patterns. Embedding-based similarity analysis further indicated that most candidate forms remained semantically close to controlled-vocabulary concepts, while only a small subset showed characteristics consistent with hallucinations. The comparison between prompting strategies revealed differences in lexical variability. Few-shot prompting led towards repetitive documentation templates, e.g., repeating the same structure and information found within the example prompt though varied slightly by changing abbreviations or contents. This meant that the few-shot method introduced rigidity with prompt examples and reduced lexical diversity, whereas zero-shot prompting produced a higher degree of structural variability and a greater proportion of candidate non-standard forms.

From a lexicographic perspective, this approach offers a systematic way to identify emerging terminology and non-standard usages in specialized corpora. By mapping novel forms to established vocabularies, lexicographers can better document language innovation, improve term standardization, and support the creation of more comprehensive, up-to-date lexical resources for clinical, technical and scientific domains.

References

Brender, T. D., Celi, L. A., & Cobert, J. M. (2025). Clinical Notes as Narratives: Implications for Large Language Models in Healthcare. *Journal of General Internal Medicine*, 40(3), 687–689. https://doi.org/10.1007/s11606-024-09093-y

Busch, F., Hoffmann, L., Rueger, C., van Dijk, E. H., Kader, R., Ortiz-Prado, E., Makowski, M. R., Saba, L., Hadamitzky, M., Kather, J. N., Truhn, D., Cuocolo, R., Adams, L. C., & Bressem, K. K. (2025). Current applications and challenges in large language models for patient care: A systematic review. *Communications Medicine*, 5(1), 26. https://doi.org/10.1038/s43856-024-00717-2

Zheng, J., Ritter, A., & Xu, W. (2024). NEO-BENCH: Evaluating Robustness of Large Language Models with Neologisms. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers) (pp. 13885–13906). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.749

Contact information

Amila Kugic
Medical University of Graz
amila.kugic@medunigraz.at

Nina Ledinek, Mija Michelizza, Špela Petric Žižić,
Domen Krvina, Janoš Ježovnik & Andrej Perdih

USING LLMS IN DICTIONARY COMPILATION

The Case of Classifying Adjectives in the Dictionary of the Slovenian Standard Language eSSKJ

Keywords Lexicography; classifying adjectives; prompt engineering; LLM; Slovene

In recent years, several lexicographic studies have addressed the question of which lexicographic processes can effectively make use of LLMs (De Schryver, 2023; Lew, 2024; McKean & Fitzgerald, 2024). Some studies are dedicated specifically to Slovenian lexicography (Gantar, 2024; Arhar Holdt et al., 2025), including Meterc & Jakop (2025) who have examined the use of LLMs in the compilation of the Dictionary of the Slovenian Standard Language eSSKJ (2016–), which is being compiled at the Fran Ramovš Institute of the Slovenian Language, ZRC SAZU, and published on the Fran dictionary portal (Ahačič et al., 2015).

This paper focuses on the performance of ChatGPT 5.5 in compiling eSSKJ dictionary entries for classifying adjectives. According to the editorial principles used in eSSKJ, classifying adjectives tend to have relatively uniform type-based definitions and predictable semantic groups of collocates, which makes them a suitable case study for testing the extent to which LLMs can successfully generate complete dictionary entries. The study tests the LLM's performance in the compilation of 35 dictionary entries for classifying adjectives derived from nouns denoting sports, academic disciplines, and animals (e.g. *košarkarski* 'pertaining to basketball', *geološki* 'pertaining to geology', *kengurujev* 'pertaining to kangaroo'), selected because the corresponding base nouns are often monosemous or exhibit regular polysemy.

The LLM was provided with the following background knowledge prior to prompting: (1) a typology of definitions used for classifying adjectives, (2) a training set of 72 previously published eSSKJ dictionary entries for classifying adjectives, (3) corpus data in the form of word sketches for these adjectives, and (4) up to 1,000 randomly selected concordances for each adjective from the Gigafida 2.0 Slovenian reference corpus (where fewer concordances were available, all examples were provided). Apart from the typology of definitions, which was supplied to the LLM in textual form, all other data were encoded in XML. The same types of data had originally been available to human editors during the compilation of the dictionary entries.

We evaluated the performance of the LLM in the following editorial tasks: (1) sense differentiation, (2) formulation of (type-based) dictionary definitions, (3) selection of typical collocations and their assignment to the appropriate dictionary senses, and (4) illustration of senses with typical dictionary examples drawn from the Gigafida 2.0 corpus. As a control measure, the same dictionary entries were also edited manually in parallel and were not uploaded to the LLM.

Testing the performance of the LLM in its editorial treatment of classifying adjectives in eSSKJ shows that it formulates type-based dictionary definitions in accordance with the established typology, although the ordering of predictable metonymic senses does not always follow the established hierarchy. It also selects appropriate collocates, both in terms of the range of collocates included and their assignment to dictionary senses, and successfully adheres to formatting guidelines for collocations (semantic clusters, formal and alphabetical ordering, etc.).

The model also performs well in the selection of dictionary examples. In these respects, the LLM's editorial treatment is comparable to dictionary entries produced by human editors. The LLM is also successful in distinguishing senses in dictionary entries that contain only predictable, type-based senses; however, it often fails to identify additional metaphorical or derived senses, as well as fixed expressions.

The results suggest that although ChatGPT 5.5 does not yet perform all tasks involved in compiling entries for classifying adjectives in eSSKJ entirely satisfactorily, it nonetheless represents a highly useful tool for lexicographers. At present, it can assist eSSKJ editors particularly with more routine tasks, which is especially valuable because these tend to be very time-consuming.

References

Ahačič, K., Ledinek, N., & Perdih, A. (2015). Fran: the next generation Slovenian dictionary portal. In K. Gajdošová & A. Žáková (Eds.), *Natural language processing, corpus linguistics, lexicography: proceedings, eighth international conference* (pp. 9-16). RAM-Verlag.

Arhar Holdt, Š., Gapsa, M., Gantar, P. & Kosem, I. (2025). Potencial ChatGPT pri razvoju Slovarja sopomenk sodobne slovenščine. *Prispevki za novejšo zgodovino*, 65(3), 189-217. https://doi.org/10.51663/pnz.65.3.08

Dictionary of the Slovenian Standard Language eSSKJ (2016–). ZRC SAZU, Založba ZRC. https://doi.org/10.3986/2738-6643

Gantar, P. (2024). Formulisanje rečničkih definicija pomoću veštačke inteligencije na primeru slovenačkih frazeoloških jedinica. In S. Marjanović (Ed.), *Leksikografski susreti 1: moderni rečnici u funkciji prosečnoga korisnika: stari problemi, savremeni pravci i novi izazovi* (pp. 151-157). Faculty of Philology, University of Belgrade. https://doi.org/10.18485/lexicog_meet.2024.1.ch12

Lew, R. (2024). Dictionaries and lexicography in the AI era. *Humanities and Social Sciences Communications*, 11, 426. https://doi.org/10.1057/s41599-024-02889-7

McKean, E., & Fitzgerald, W. (2024). The ROI of AI in Lexicography. *Lexicography. Journal of Asialex,* 11(1). https://doi.org/10.1558/lexi.27569

Meterc, M., & Jakop, N. (2025). Identifying the most representative phraseological units using language corpora and artificial intelligence for lexicography: the case of Slovenian comparative phrasemes. In I. Kosem et al. (Eds.), *Electronic lexicography in the 21st century (eLex 2025): intelligent lexicography, Proceedings of the eLex 2025 Conference* (pp. 795-814). https://elex.link/elex2025/wp-content/uploads/eLex2025-46-MetercJakop.pdf

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

Contact information

Nina Ledinek
ZRC SAZU, Fran Ramovš Institute of the Slovenian Language
nina.ledinek@zrc-sazu.si

Mija Michelizza
ZRC SAZU, Fran Ramovš Institute of the Slovenian Language
mija.michelizza@zrc-sazu.si

Špela Petric Žižić
ZRC SAZU, Fran Ramovš Institute of the Slovenian Language
spela.petric@zrc-sazu.si

Domen Krvina
ZRC SAZU, Fran Ramovš Institute of the Slovenian Language
domen.krvina@zrc-sazu.si

Janoš Ježovnik
ZRC SAZU, Fran Ramovš Institute of the Slovenian Language
janos.jezovnik@zrc-sazu.si

Andrej Perdih
ZRC SAZU, Fran Ramovš Institute of the Slovenian Language
andrej.perdih@zrc-sazu.si

Acknowledgements

This article was written within the Slovenian Research Agency's P6-0038 The Slovenian Language in Synchronic and Diachronic Development program group.

Seulki Lee

REFRAMING WAN (完), HAO (好), AND DAO (到) IN CHINESE–KOREAN BILINGUAL LEXICOGRAPHY

A Corpus-Based Cognitive Analysis

Keywords Chinese resultative complements; bilingual lexicography; image schema; Chinese–Korean parallel corpus; exhaustion; readiness; attainment

Introduction

Chinese resultative complements such as *Wan* (完), *Hao* (好), and *Dao* (到) are frequently treated in bilingual dictionaries as generalized markers of completion or result. However, existing lexicographic descriptions often rely on simplified lexical glosses such as 'finish', 'good', or 'arrive', obscuring important cognitive-functional distinctions among these complements. In actual usage, the three markers exhibit differentiated event-structural patterns, semantic preferences, and translational correspondences. This study reexamines *Wan*, *Hao*, and *Dao* through a Chinese–Korean parallel corpus and argues for a cognitively informed, corpus-based approach to bilingual lexicography, with implications for learner-oriented and AI-assisted dictionary design.

Theoretical framework

Following corpus-driven approaches to bilingual lexicography (Atkins & Rundell, 2008), this study adopts a cognitive-functional perspective combining cognitive scanning and image-schema theory. Langacker's notion of cognitive scanning explains how speakers construe events (Langacker, 1987, pp. 244–249), while Johnson's image schemas, such as CONTAINMENT and PATH (Johnson, 1987, pp. 21–30), provide a framework for modeling boundedness, resultant configuration, and goal attainment.
Figure 1 presents the cognitive-image-schematic profiles of the three complements. *Wan* profiles a bounded process brought to terminal exhaustion, *Hao* profiles the transformation of a disordered or incomplete state into an acceptable configuration, and *Dao* profiles movement or access toward a spatial, perceptual, or abstract target.

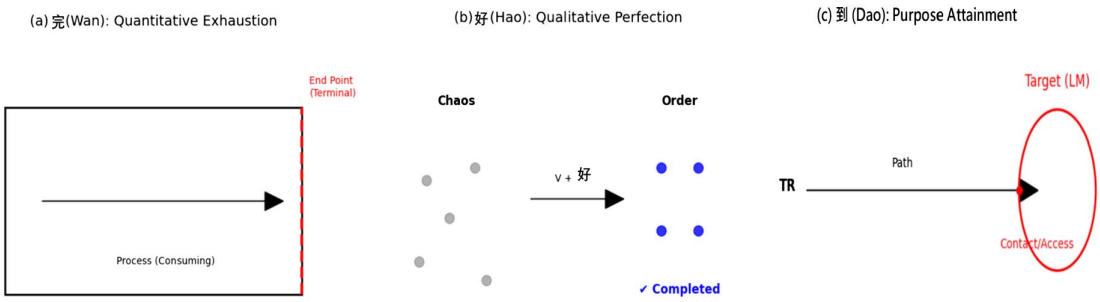


FIGURE 1 Image-schematic profiles of Wan (完), Hao (好), and Dao (到)

Methodology

The dataset consists of 871 valid instances of *Wan*, *Hao*, and *Dao* extracted from the AI Hub Chinese–Korean parallel corpus of broadcast discourse. Each example was manually annotated for verb class, verb supercategory, object type, object-role structure, functional group, aspectual subtype, and Korean translational correspondence. Frequency distributions, percentage comparisons, and χ^2 tests were used to identify statistically differentiated semantic-functional tendencies.

Findings

The corpus analysis reveals that *Wan* (完), *Hao* (好), and *Dao* (到) form statistically differentiated cognitive-functional profiles across verb classes, object preferences, functional distributions, and Korean translational correspondences.

First, the three complements exhibit significantly different verb-class distributions (χ^2, $p < .001$). *Wan* preferentially co-occurs with consumption, communication, and creation-related verbs, reflecting exhaustive completion and resource depletion. Patterns such as 吃完, 用完, and 做完 profile bounded completion and sequential event closure. In contrast, *Hao* frequently occurs with action, arrangement, and cognition-related verbs, indicating successful configuration or readiness, as in 准备好 and 安排好. *Dao* shows strong associations with perception, cognition, and motion verbs, including 看到, 听到, and 找到, reflecting attainment toward perceptual, cognitive, or goal-oriented endpoints.

Second, functional-group distributions also differ significantly (χ^2, $p < .001$). *Wan* is associated with SEQUENTIAL and CONFIGURATION functions, and also appears in MODALITY-related contexts involving obligation release, such as 做完就不用 . . . . *Hao* is strongly associated with DIRECTIVE and STATUS_NOTIFICATION functions, signaling preparation, readiness, or resultant-state confirmation. *Dao* is linked to EVALUATION and STATUS functions, highlighting attainment and realization of abstract or perceptual targets.

Third, object-type distributions reinforce these distinctions. *Wan* tends to co-occur with bounded eventive and discourse-related objects, including tasks, speech content, and information. Hao predominantly combines with manipulable concrete entities, reflecting state-configuration processes. *Dao* frequently occurs with abstract, discourse-related, and locational objects, emphasizing endpoint-oriented attainment.

Finally, Korean translational correspondences also reflect these profiles. *Wan* is often translated as '-다', '끝내다', or '다 해버리다'; *Hao* as '준비되다', '잘', or '해 놓다'; and

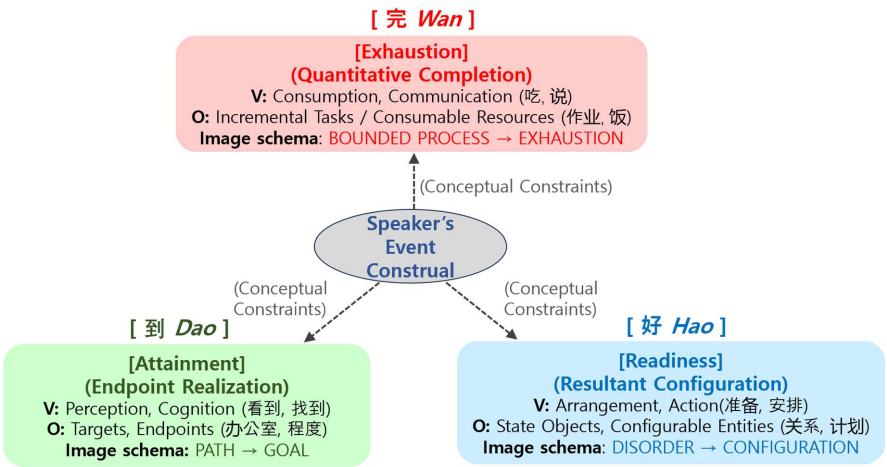


FIGURE 2 Integrated [V-RC-O] selection-constraint model for *Wan*, *Hao*, and *Dao*

TABLE 1 Dominant semantic profiles of *Wan, Hao*, and *Dao*

Complement	Dominant verb classes	Functional tendency	Korean mapping tendency
完 (*Wan*)	consumption / communication	exhaustive completion / sequential boundary	다 / -고나서
好 (*Hao*)	arrangement / action	readiness / configuration	준비되다 / 잘 / -해놓다
到 (*Dao*)	perception / cognition	attainment / evaluation	찾다 / 도달하다

Dao as '도달하다', '찾다', '보다', or '듣다'. These patterns show that the semantic distinctions among Wan, Hao, and Dao are reflected in both Chinese corpus distributions and bilingual translational behavior.

Based on these distributional patterns, the study proposes an integrated [V-RC-O] selection-constraint model in which the three complements interact differently with verb semantics, object boundedness, and image-schematic patterns.

Lexicographic implications and conclusion

These findings suggest that bilingual dictionaries should not reduce *Wan*, *Hao*, and *Dao* to isolated glosses such as 'finish', 'good', or 'arrive'. Instead, entries should reflect their distinct cognitive-functional profiles: *Wan* as exhaustive completion, *Hao* as readiness-oriented configuration, and *Dao* as attainment-oriented realization. By integrating corpus distributions, functional tendencies, and Korean translational mappings, this study proposes a cognitively informed, corpus-based model for Chinese–Korean bilingual lexicography.

References

AI Hub. (n.d.). Korean–Chinese Parallel Corpus. Retrieved May 2026, from https://aihub.or.kr

Atkins, S. B. T. & Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford University Press.

Langacker, R. W. (1987). *Foundations of Cognitive Grammar Volume I: Theoretical Prerequisites*. Stanford University Press.

Johnson, M. (1987). *The Body in the Mind: The Bodily Basis of Meaning, Imagination, and Reason*. University of Chicago Press.

Contact information

Seulki Lee
Chungbuk National University
nono0118@naver.com

Magdalena Majdak, Ewa Rodek, Katarzyna Kryńska & Jagoda Marszałek

THE ROLE OF LLMS IN THE AUTOMATIC DISTRIBUTION OF MEANINGS FROM KNAPIUSZ'S POLISH–LATIN–GREEK DICTIONARY OF 1643

A Pilot Study

Keywords Historical lexicography; LLMs; sense disambiguation; Polish lexicography; multilingual dictionaries

Introduction

Grzegorz Knapiusz's *Thesaurus polonolatinograecus* (second, expanded edition 1643; hereafter Kn) is the largest seventeenth-century dictionary that also records Polish vocabulary. It is one of the source materials for the *Electronic Dictionary of Seventeenth- and Eighteenth-Century Polish* (hereafter e-SXVII). One challenge for e-SXVII editors is assigning the content of Knapiusz's entries—which contain, in addition to Polish, extensive Latin and Greek material provided as equivalents of Polish items—to the senses distinguished in e-SXVII entries. Although the *Thesaurus* was intended primarily as an extensive synonym dictionary supporting Latin learning, it also contains numerous observations on Polish vocabulary. It is regarded as the first work in Polish lexicography to address the semantics of Polish words. Knapiusz does not explicitly divide meanings within entries; instead, the Latin and Greek equivalents, or associated lexical items, of a Polish headword shift fluidly from one meaning to another, without clearly marked boundaries.

Aims

This study aimed to test whether LLMs can support the segregation and disambiguation of meanings in Kn entries, as well as their assignment to corresponding senses in e-SXVII. We designed and evaluated a procedure for the semi-automatic distribution of Kn material within e-SXVII entries using GPT-5.2 Thinking, Gemini 3 Pro Thinking, Perplexity Sonar Pro, and Claude Sonnet 4.5.

Method

We tested ten polysemous nominal entries, including *mieczyk* 'short sword; gladiolus; yellow iris; bulbous iris; stinking iris; swordfish; Prussian coin', *dowód* 'proof; document; evidence; demonstration; reason', *laska* 'staff; mace; wand; rod; ridge; office; shepherd's purse; unit of measure', *wyrok* 'oracle; decree; sentence; fate', *zapis* 'record; deed; bequest; legal act; treaty; statute; register; address; lodging assignment', *gwałt* 'assault; force; outrage; rape; corvée labour; violence; necessity; tumult', *wynalazek* 'idea; invention; discovery; decision; finding; result; decree; compromise', *bukowina* 'beech forest; woodland; beech trees; beech wood; beech grove; forest settlement', *motłoch* 'rabble; mob; worthless mass; commoners; riffraff', and *jabłko* 'apple; apple tree; apple-like fruit; pomegranate; truffle; pomander; heraldic charge'.

The trials differed in the amount and type of information provided to the models: (1) a basic instruction to translate, segregate and disambiguate Latin and Greek equivalents in Kn entries; (2) the same instruction with reference to the sense structure and definitions of model e-SXVII entries; (3) the instruction from trial 2 supplemented with quotations and established word combinations; (4) the instruction from trial 2 supplemented with the e-SXVII principles for defining senses and distinguishing polysemy from homonymy; and (5) the instruction from trial 4 supplemented with illustrative material from e-SXVII.

To reduce the influence of conversational context, each Kn entry was queried separately, in a new conversation. In evaluating the results, we took into account inter-entry variability resulting, among other factors, from different degrees of lexical polysemy and different entry structures. Outputs were compared with model e-SXVII entries and evaluated by lexicographers and classical philologists according to an agreed scale covering the correctness of sense division, the assignment of equivalents to senses, and agreement with e-SXVII definitions.

Results and conclusions

Perplexity Sonar Pro was excluded because its responses were not consistently generated by the same model. GPT-5.2 Thinking, Gemini 3 Pro Thinking, and Claude Sonnet 4.5 produced outputs of highly variable accuracy. Each model performed flawlessly on some entries but made significant errors in others. This variability appears to be related to differences in polysemy and entry structure. Increasing prompt specificity did not produce a straightforward improvement in accuracy, suggesting that output instability cannot be resolved merely by enlarging the sample without changing the validation procedures.

The trials also examined the minimum amount of information needed for an AI model to identify the semantic structure of a Kn entry correctly. For entries with highly typical mechanisms of polysemy, such as *bukowina* ‘a forest dominated by beech trees’ → ‘individual beech trees’ → ‘beech wood’, satisfactory results were obtained as early as the first trial. However, in comparable cases, such as *jabłko* ‘apple fruit’ → ‘apple tree’ or ‘round fruit’ → ‘round object’, the results were unsatisfactory both in the first trial and after additional context had been supplied. For some entries, subsequent trials even led to a deterioration in the disambiguation of Kn meanings, regardless of the AI model used.

The pilot study revealed clear limitations of the tested procedure. Verification of the method showed that any attempt to automatically divide meanings within loosely structured historical dictionary entries must take into account both their content and their intended function. In the case of Knapiusz’s *Thesaurus*, most illustrative material was taken from ancient authors and may therefore represent meanings not attested in Polish. Moreover, the study indicates that the current functioning of the selected models does not allow for standardised results or fully automatic support for lexicographic work. Nevertheless, under appropriate methodological assumptions, AI models with a “thinking” mode may offer support to lexicographers by generating hypotheses about senses, identifying problematic equivalents, and comparing Kn material with e-SXVII definitions.

References

W. Gruszczyński (Ed., 2004–2023), M. Majdak (Ed., 2024–). e-SXVII: *Elektroniczny słownik języka polskiego XVII i XVIII wieku*. (n.d.). Retrieved May 24, 2026, from https://sxvii.pl

Knapiusz, G. (1643). *Thesaurus Polonolatinograecus seu promptuarium linguae Latinae et Graecae* [...] (Vol. 1, 2nd ed.). Drukarnia Łazarzowa.

Contact information

Magdalena Majdak
Polish Academy of Sciences
magdalena.majdak@ijppan.pl

Ewa Rodek
Polish Academy of Sciences
ewa.rodek@ijppan.pl

Katarzyna Kryńska
Polish Academy of Sciences
katarzyna.krynska@ijppan.pl

Jagoda Marszałek
Polish Academy of Sciences
jagoda.marszalek@ijppan.pl

Acknowledgements

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.

Leonor Martins & Ana Salgado

NEOLOGISMS IN EVARISTO.AI *VS.* ChatGPT

Implications for LLM-Based Lexicography in European Portuguese

Keywords LLM-based lexicography; neologisms; Evaristo.ai; ChatGPT-5.2; European Portuguese

The diffusion of large language models (LLMs) has generated growing interest in their application to lexicographic practice, particularly for tasks such as the automatic generation of dictionary entries. Existing studies generally report promising results; however, they have primarily focused on major languages, most notably English, and rely on international, multilingual LLMs, such as ChatGPT (De Schryver, 2023; Lew, 2023; Rees & Lew, 2023), thereby overlooking the performance of language-specific, native LLMs. Accordingly, the lexicographic treatment of neologisms by LLMs has also received little attention. As emerging and often unstable lexical units that are sparsely attested or entirely absent from training data, neologisms limit empirical evidence for reliable semantic generalization and force models to rely on internal mechanisms of meaning composition rather than memorized patterns. This raises specific concerns regarding the extent to which LLMs can model lexical innovation in ways that are consistent with established lexicographic standards (Poix & Shevchenko, 2025). Some of the studies on LLMs and neologisms have approached the issue from a morphological perspective. Studies conducted in English and Polish (Manova, 2023) and Greek (Georgiou, 2025) suggest that LLMs perform unevenly across neologisms derived from different word-formation processes, generally achieving better processing results for derivatives than for compounds. However, since languages differ in their morphology and word-formation processes, it remains unclear to what extent these observations extend to other languages. In this context, the present study presents a comparative analysis of Evaristo.ai, the first chatbot designed for the Portuguese language, and ChatGPT-5.2, a multilingual LLM, focusing on their ability to generate lexicographic entries for neologisms arising from different word-formation processes, namely compounds, derivatives, and loanwords. Specifically, it examines whether and how word-formation influences the automatic lexicographic treatment of neologisms, assessing whether the contrasts previously observed between derivatives and compounds are reproduced in generated dictionary entries, and the extent to which the resulting entries adhere to established lexicographic standards.

This analysis was based on a set of fifteen neological units extracted from the Observatório Lexical (OL), a repository dedicated to the monitoring of lexical units that have not yet been incorporated into the Dicionário da Língua Portuguesa (DLP) of the Academy of Sciences of Lisbon, but whose attestation in contemporary usage warrants linguistic observation. These neologisms were subsequently categorized according to their word-formation process (Correia & Payo de Lemos, 2005; Rio-Torto et al., 2013) into (i) compounds (ciberautoritarismo, criptocorrupção, lgbtfóbico, neonepotismo, tecnomelancólico), (ii) derivatives (almirantismo, bolhificação, lepenismo, urgencializar, trumpização) and (iii) loanwords (brain rot, bromance, doomscrolling, gamekeeper, vlogger). A set of prototype dictionary entries was manually constructed in accordance with the DLP's criteria and used as a gold standard. Subsequently, all neologisms were processed using a single standardised prompt (Lo, 2023), which specified the required entry structure and the mandatory lexicographic fields. The generated entries were evaluated by two annotators according to a systematic methodology based on the

lexicographic conventions of the DLP, considering dimensions such as structural compliance, lemma presentation, grammatical category, definition, example of use, etymology, variety sensitivity (pt-PT), lexicographic style and register, and error typology. Each dimension was rated on a three-point scale (0–2), and entries were then classified as inadequate, partially adequate, or adequate, according to the deficiencies in core lexicographic fields and the degree of editorial intervention required. The results indicate that ChatGPT-5.2 generates lexicographic entries of higher overall adequacy across all neologism types, an expected outcome plausibly related to its training on substantially larger volumes of data. More importantly, however, the distribution of performance across word-formation processes diverges from the patterns reported in previous studies. In the present dataset, lexicographic entries for derivatives do not outperform those for compounds and instead display substantially lower adequacy, a pattern that is especially pronounced for Evaristo.ai: among the derivatives, Evaristo.ai produced three inadequate definitions, one partially adequate definition, and one adequate definition, whereas for compounds it produced two adequate definitions and three partially adequate definitions. This underperformance is primarily attributable to flawed etymological analyses, often involving implausible, hallucinated historical details or misidentified source languages or formative elements, as well as to deficiencies in the example field, which is frequently unclear, implausible, or poorly integrated with the lemma. These findings suggest that the automatic lexicographic treatment of neologisms by LLMs in European Portuguese may be influenced by word-formation processes, though not in the manner previously reported for other languages. Contrary to earlier findings, derivational neologisms appear to be more challenging than compounds, particularly for the language-specific model Evaristo.ai. This indicates that language-specific training does not systematically mitigate previously observed weaknesses. Overall, the study contributes to LLM-based lexicography by showing that word-formation processes affect not only neologism recognition, but also the quality and editorial reliability of automatically generated dictionary entries for European Portuguese.

References

Correia, M., & Payo de Lemos, L. S. (2005). *Inovação lexical em português*. Edições Colibri e Associação de Professores de Português.

Dardano, M., Frenguelli, G., & Colella, G. (2006). What Lexicographers Do with Word Formation. In E. Corino, C. Marello & C. Onesti (Eds.), *Proceedings of the XII EURALEX International Congress: Lexicological Issues of Lexicographical Relevance* (pp. 1115-1127). Edizioni dell'Orso.

De Schryver, G. M. (2023). Generative AI and Lexicography: The Current State of the Art Using ChatGPT. *International Journal of Lexicography*, 36. ecad021. https://doi.org/10.1093/ijl/ecad021

DLP: *Dicionário da Língua Portuguesa da Academia das Ciências de Lisboa*. Retrieved April 6, 2026, from https://dicionario.acad-ciencias.pt/

Freixa, J. (2022). El proceso neológico. In J. Freixa, S. Torner & E. Bernal (Eds.), *La neología del español: del uso al diccionario* (pp. 51-72). Iberoamericana Vervuert.

Georgiou, G. (2025). Can AI mimic the human ability to define neologisms? *International Journal of Speech Technology*, 28, 773–782. https://doi.org/10.1007/s10772-025-10216-x

Lew, R. (2023). ChatGPT as a COBUILD lexicographer. *Humanities and Social Sciences Communications* 10, 704. https://doi.org/10.1057/s41599-023-02119-6

Lo, L. S. (2023). The Art and Science of Prompt Engineering: A New Literacy in the Information Age. *Internet Reference Services Quarterly*, 1-8. https://doi.org/10.1080/10875301.2023.2227621

Manova, S. (2023). ChatGPT, n-grams and the power of subword units: The future of research in morphology. In M. Filko & K. Šojat (Eds.), *Proceedings of the Fourth International Workshop on Resources and Tools for Derivational Morphology: DeriMo 2023* (pp. 1-11). Hrvatsko društvo za jezične tehnologije.

OpenAI. ChatGPT-5.2. Retrieved April 6, 2026, from https://chatgpt.com/

Poix, C. & Shevchenko, N. (2025). The challenge of AI-generated neology. In I. Kosem, M. Jakubíček, M. Medveď, K. Zgaga, Š. Arhar Holdt, T. Munda & Salgado, A. (Eds.), *Electronic lexicography in the 21st century (eLex 2025): Intelligent lexicography. Proceedings of the eLex 2025 conference* (pp.318-331). Lexical Computing CZ, s.r.o.

PORTULAN CLARIN: *Evaristo.ai*. Retrieved April 6, 2026, from https://evaristo.ai/

Rees, G. P. & Lew, R. (2023). 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, 50–74. https://doi.org/10.1093/ijl/ecad030

Rio-Torto, G., Rodrigues, A., Pereira, I., Pereira, R. & Ribeiro, S. (2013). *Gramática Derivacional do Português*. Imprensa da Universidade de Coimbra.

Contact information

Leonor Martins
Institute of Lexicology and Lexicography of the Portuguese Language, Academy of Sciences of Lisbon
lmartins@acad-ciencias.pt

Ana Salgado
University of Porto & Academy of Sciences of Lisbon
acsalgado@letras.up.pt

Acknowledgements

This work is supported by: (1) PORTULAN CLARIN – Research Infrastructure for Language Science and Technology [LISBOA2030-FEDER-01316900], funded by Programa LISBOA2030 – Programa Regional de Lisboa 2021–2027; and (2) Centre for Linguistics of the University of Porto (CLUP), funded by national funds through FCT – Fundação para a Ciência e a Tecnologia, I.P., under project UID/00022/2025 (https://doi.org/10.54499/UID/00022/2025).

Abdelouahab Masbah, Andrea Bellandi & Silvia Piccini

DEC FACILE

Semantic-Web Authoring for Combinatorial Explanatory Dictionaries

Keywords Explanatory Combinatorial Dictionary; Ontolex-Lemon; semantic web; linked data

This paper presents DEC Facile (Bellandi et al., 2025), a tool for creating Explanatory Combinatorial Dictionaries (ECDs) in a Semantic Web environment. Grounded in Explanatory Combinatorial Lexicology (Mel'čuk et al., 1995) and adopted in both general and specialized lexicography, ECDs provide a highly structured representation of lexical meaning and combinatorial properties.

Because of their expressiveness, however, ECDs are complex and time-consuming to build. Earlier ECD editors (Sérasset, 1996; Gader et al., 2012) supported the formalization of lexicographic data; DEC Facile extends this line of work by enabling the creation of machine-actionable and interoperable ECDs compatible with Semantic Web standards. Its design is based on three principles: ease of use, especially for humanists unfamiliar with RDF/OWL; support for collaborative lexicographic work; and the production of FAIR lexical resources (Wilkinson et al., 2016).

The system[1] is built on LexO-server (Bellandi, 2023), a Java-based backend exposing RESTful services for dictionary and sense management. It orchestrates RDF models based on OntoLex-Lemon[2] and its extensions, notably Lexicog, SynSem, LexFom (Fonseca et al., 2016), and SKOS. Lexical data are stored in GraphDB and accessed through SPARQL queries. The graphical interface is a responsive single-page web application developed in TypeScript with Angular, PrimeNG, Tailwind CSS, and PrimeIcons. It supports structured editing, dynamic language switching, and real-time interaction with backend services.

The main interface (Figure 1) offers an integrated workspace for navigating, creating, and managing lexical resources through a modular three-panel layout that hides the complexity of Semantic Web technologies from non-technical users. The left-hand panel provides access to the dictionary inventory, with functions for creating, importing, exporting, searching, and filtering dictionaries. The central panel organizes dictionary settings into thematic sections, while the right-hand panel contains the dictionary creation form, where users define core metadata such as label, language, and description. This design supports multilingual projects, collaborative workflows, and immediate synchronization with RDF data stored in GraphDB.

Dictionary entries are listed with essential information (lemma, part of speech, and language), and can be expanded to reveal morphological forms and lexical senses. For each entry, a set of action icons allows users to switch between consultation and editing modes, manage editorial status (e.g. working, completed, reviewed), or perform structural operations such as deletion. The interface also allows lexicographers to create new dictionary entries directly from this panel, facilitating continuous lexicographic work.

[1] DEC Facile is released as an open-source project; the code and documentation are accessible via Github at the following link: https://github.com/Abdoumasbah/DEC-Facile

[2] https://www.w3.org/2016/05/ontolex/

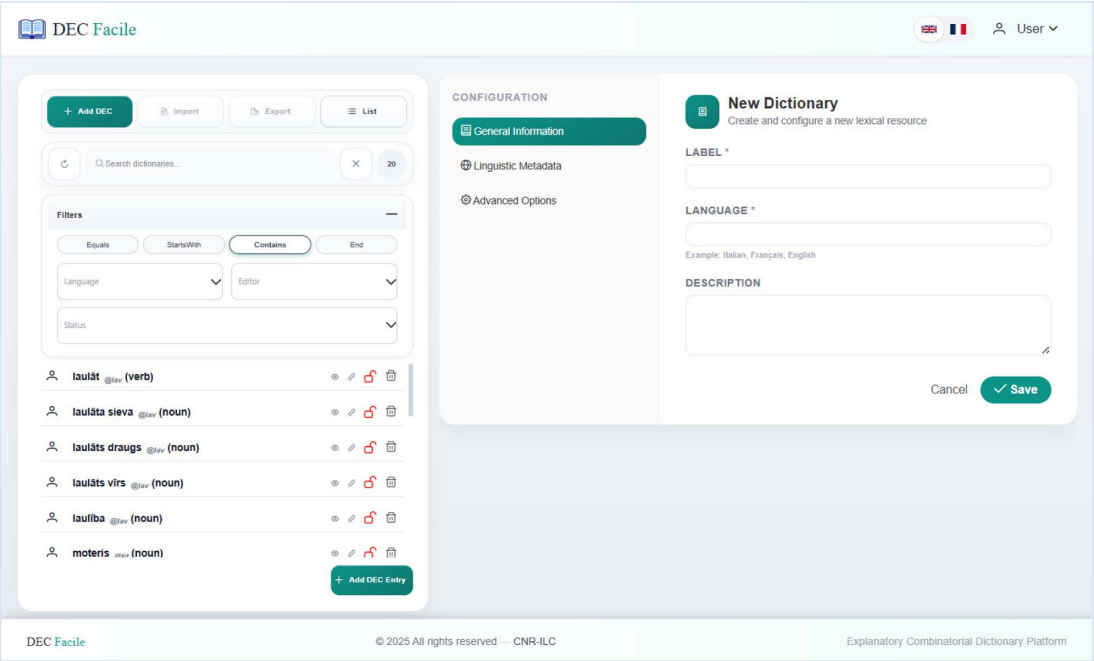


FIGURE 1 Creation of a new dictionary in DEC Facile

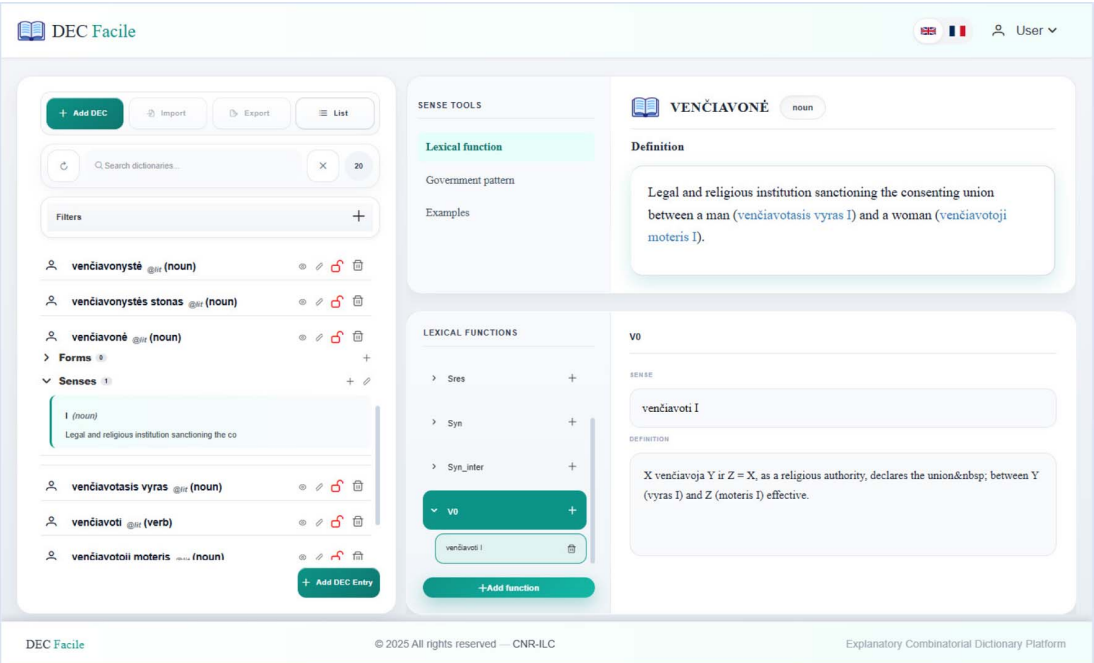


FIGURE 2 Editing a dictionary entry in DEC Facile: definitions and lexical functions

When a dictionary entry is selected from the left-hand panel (Figure 2), the central workspace provides section-based access to its main components, i.e. definitions, government pattern, lexical functions, and usage examples.

DEC Facile supports the linking of lexical senses within definitions: during editing, lexicographers can select an existing sense from a controlled autocompleted list and insert it directly into the definition. As shown in Figure 2, the definition of the entry *venčiavonė* contains

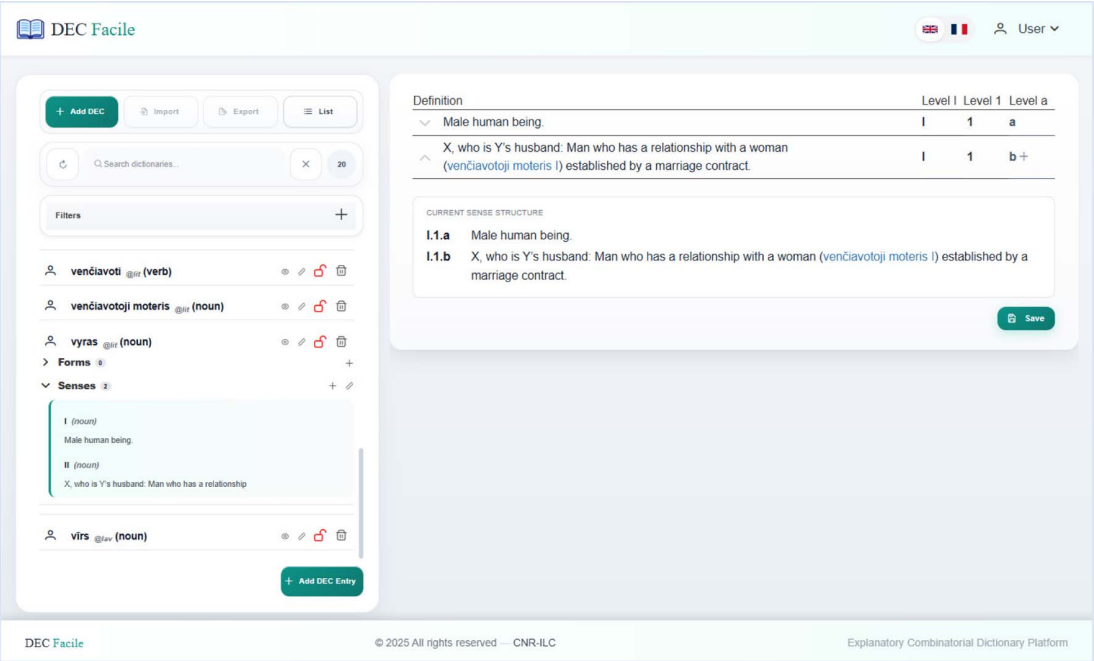


FIGURE 3 Sense hierarchy management in DEC Facile.

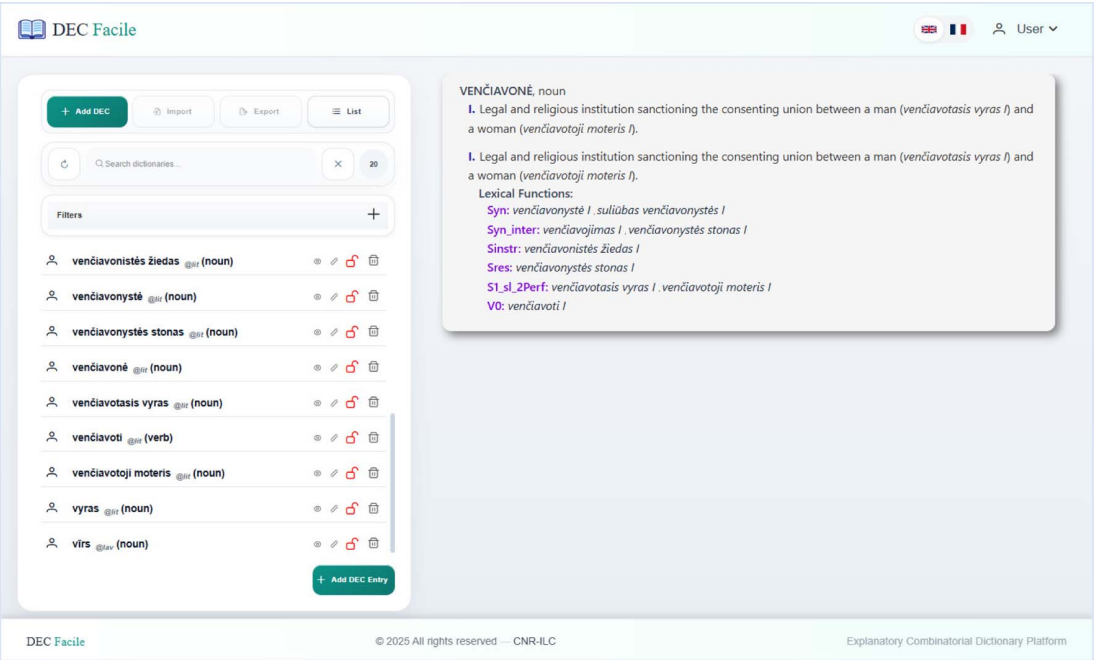


FIGURE 4 Dictionary view in DEC Facile

cross-references to other lexical senses already encoded in the dictionary; in the RDF model, these links are represented through rdfs:seeAlso. This mechanism reduces inconsistencies and encoding errors, while enabling the computational reuse of semantic relations and providing immediate visual feedback through a definition preview.

Lexical functions are instantiated by linking target lexical senses already defined in the dictionary, rather than by manually entering strings. Lexical relations are therefore represented

at sense level, improving semantic coherence and interoperability. A dedicated sense-management interface (Figure 3) supports the hierarchical organization of senses according to graded semantic distance, in line with ECD conventions. Lexicographers can assign levels, adjust ordering, and visualize the structure in real time.

DEC Facile also provides a dictionary-style read-only view (Figure 4), which reproduces the layout of traditional printed ECDs and displays senses, definitions, lexical functions, and examples. The resulting ECDs are represented in RDF within the OntoLex-Lemon ecosystem, stored in GraphDB, and exportable to LLOD endpoints. The tool is currently used in *Old words for a New World*[3] and the RUT project.

References

Bellandi, A. (2023). Building linked lexicography applications with LexO-server. *Digital Scholarship in the Humanities*, *38*(3), 937-952. https://doi.org/10.1093/llc/fqac095

Bellandi, A., Masbah, A., & Piccini, S. (2025). DEC Facile: Bringing Combinatory Explanatory Dictionaries to the Semantic Web. *Journal of Digital Terminology and Lexicography*, 1(JDTL-Volume 1, issue 2), 81-98.

Fonseca, A., Sadat, F., & Lareau, F. (2016). Lexfom: A lexical functions ontology model. In M. Zock, A. Lenci & S. Evert (Eds.), *Proceedings of the 5th Workshop on Cognitive Aspects of the Lexicon (CogALex - V)* (pp. 145-155). The COLING 2016 Organizing Committee.

Gader, N., Lux-Pogodalla, V., & Polguère, A. (2012). *Hand-crafting a lexical network with a knowledge-based graph editor*. In M. Zock & R. Rapp (Eds.), *Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon* (pp. 109-126). The COLING 2012 Organizing Committee.

Mel'čuk, I. A., Clas, A. P., & Polguère, A. (1995). *Introduction à la lexicologie explicative et combinatoire*. Duculot.

Sérasset, G. (1996). Un éditeur pour le dictionnaire explicatif et combinatoire du français contemporain. *Journées lexique du PRC-CHM*, 131-138. HAL.

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., & Bourne, P. E. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3(1), 1-9.

Contact information

Abdelouahab Masbah
Istituto di Linguistica Computazionale "A. Zampolli"
masbah.abdelouahab@gmail.com

Andrea Bellandi
Istituto di Linguistica Computazionale "A. Zampolli"
andrea.bellandi@ilc.cnr.it

Silvia Piccini
Istituto di Linguistica Computazionale "A. Zampolli"
silvia.piccini@ilc.cnr.it

[3] The project is funded by the Italian Ministry of University and Research: https://ownwproject.wordpress.com/

Nikos Mathioudakis & Alexandra Fiotaki

DOCUMENTING THE UNDOCUMENTED

Leveraging LLMs for Neologism Integration into Lexicons

Keywords Neologisms; undictionaried words; Large Language Models (LLMs); AI-assisted lexicography; computational lexicography; journalistic language; Modern Greek linguistics

In the present study, we aim to analyze and interpret neologisms appearing in the news streams of Greek newspapers by leveraging Large Language Models (LLMs), thereby reinforcing the notion of a bidirectional relationship between journalism and neology, while highlighting the role of Artificial Intelligence in lexicological analysis and contemporary lexicography. The aim is to enrich and update Modern Greek lexicons by converting attested journalistic neologisms into dictionary-ready entries, with the lexicographer retaining final authority over inclusion and editorial form.

In the digital age, journalism faces both challenges and opportunities regarding lexical choices, as the rapid evolution of technology and media influences the way news is constructed and disseminated. Accordingly, it capitalizes on the playful and inventive use of language through neology, adapting its lexical repertoire to social and technological developments. Journalists employ neologisms, loanwords, hybrid formations, and creative linguistic compounds to capture public attention. Journalistic neologisms are not merely new words that have yet to be included in Modern Greek dictionaries; they are tools that both reflect and shape reality, encapsulating complex phenomena with ease and ingenuity.

This study moves beyond simple extraction, designing and refining a criteria-driven protocol that transitions candidate neologisms through a rigorous validation pipeline toward prioritized editorial recommendation. The pipeline is conceived as three sequential stages, namely candidate detection, multi-model validation, and weighted prioritization, each with explicit inputs, decision thresholds, and documented fallback procedures, so that the route from raw newspaper text to a vetted dictionary candidate is fully traceable and reproducible.

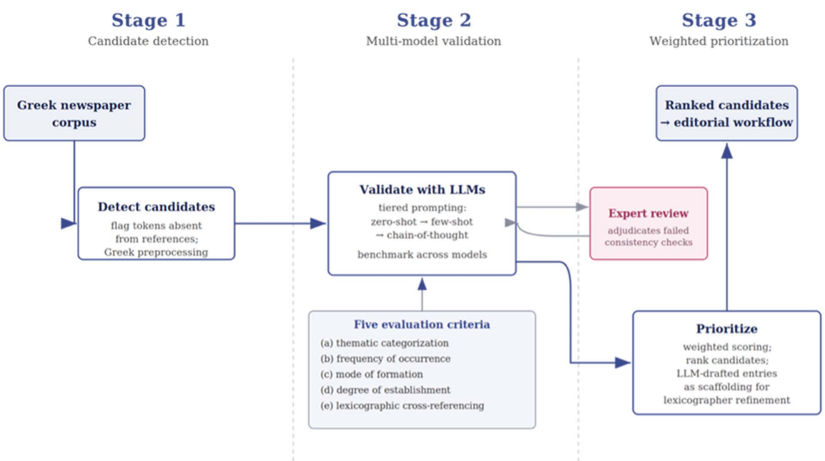


FIGURE 1 The three-stage validation-and-prioritization pipeline for documenting journalistic neologisms, from raw corpus to ranked editorial candidates

Candidate detection operates over a continuously updated corpus of Greek newspaper articles. Tokens and multiword expressions absent from reference resources are flagged as candidates, while Greek-specific preprocessing addresses morphological disambiguation, lemmatization, and orthographic normalization, ensuring that inflectional variants and accentuation differences are consolidated rather than counted as distinct items.

The validation stage deploys multiple LLMs with tiered prompt engineering, progressing from zero-shot to few-shot and chain-of-thought strategies, to apply five evaluation criteria governing progression. These comprise: (a) thematic categorization, identifying domain relevance across politics, technology, economy, society, and culture; (b) frequency of occurrence, analyzing corpus trajectory as an indicator of viability; (c) mode of formation, classifying borrowing, compounding, blending, or derivation; (d) degree of establishment, distinguishing hapax legomena for exclusion, bursty formations for monitoring, and stabilizing neologisms for advancement; and (e) lexicographic cross-referencing, confirming novelty against established dictionaries, Greek corpora, and extended attestation sources. Comparing model outputs flags disagreements for resolution, and mandatory expert review adjudicates candidates that fall below threshold on any criterion.

The prioritization stage applies weighted scoring across the validated criteria, emphasizing frequency stability, domain spread, and attestation diversity, to generate ranked candidate lists for editorial workflow integration. LLM-generated draft entries, including definitions, formation analysis, and exemplification, serve as preliminary scaffolding for lexicographer refinement rather than as finished products. These drafts are structured to lexicographic conventions (headword and lemma form, grammatical labels, etc.) so that the lexicographer's task is curation and adjudication rather than composition from scratch, keeping the methodology firmly within the lexicographic tradition it seeks to extend. Each criterion carries explicit decision thresholds and documented fallback procedures for ambiguous cases, and the weights themselves are subject to iterative refinement based on lexicographer feedback, allowing the protocol to adapt to editorial priorities over successive cycles.

Evaluation measures protocol efficacy through downstream utility: conversion rates between stages, throughput of accepted entries, reduction in editorial processing time, and alignment with existing dictionary conventions. The output is a documented, tested, and scalable procedural framework that accelerates lexical documentation while preserving essential human adjudication at every decision point.

In conclusion, the multifaceted use of neologisms in journalism, primarily manifesting in two tendencies, the creative and the descriptive, underscores the fact that neologisms constitute a significant object of study not only for linguistics and lexicography but also for communication and media/journalism studies.

References

Algeo, J. (Ed.) (1991). *Fifty years among the new words: A dictionary of neologisms, 1941–1991*. Cambridge University Press.

Anastasiadi-Symeonidi, A. (1986). *Neology in Modern Greek [Η νεολογία στην κοινή νεοελληνική]*. Aristotle University of Thessaloniki.

Christofidou, A. (2024). Neology. In G. K. Giannakis (Ed.), *Encyclopedia of Greek Language and Linguistics*. Brill.

Christofidou, A. (Ed.) (2025). *Bulletin of Scientific Terminology and Neologisms* [*Δελτίο Επιστημονικής Ορολογίας και Νεολογισμών*], iss. 16–17. Academy of Athens – Research Centre for Scientific

Terms and Neologisms [Digital edition: https://www.academyofathens.gr/el/ereyna/kentra/ereyna-neologismon/ekdoseis/psiphiakes-ekdoseis].

Georgiou, G. P. (2025). Can AI mimic the human ability to define neologisms? *International Journal of Speech Technology*, 28(4), 773–782.

Mathioudakis, N. (2026 – forthcoming). Journalism and Neology: Lexical choices in a digital age [Δημοσιογραφία και Νεολογία: λεξικές επιλογές σε μια ψηφιακή εποχή]. In N. Tsitsanoudi-Mallidi & D. Stefanou (Eds.), *The Greek language in diaspora media [Η ελληνική γλώσσα στα ομογενειακά μέσα ενημέρωσης]*. Pedio.

Nakas, T. & Gavriilidou, Z. (2005). *Journalism and Neology. Titles – Topical findings – Surprises [Δημοσιογραφία και Νεολογία. Τίτλοι – Ευρήματα για θέματα – Εκπλήξεις]*. Patakis.

Poix, C. & Shevchenko, N. (2025). The challenge of AI-generated neology. In I. Kosem, M. Jakubíček, M. Medveď, K. Zgaga, Š. Arhar Holdt, T. Munda & A. Salgado (Eds.), *Electronic lexicography in the 21st century (eLex 2025): Intelligent lexicography. Proceedings of the eLex 2025 conference* (pp. 322–335). Lexical Computing CZ.

Sablayrolles, J.-F. (1997). Néologismes: une typologie des typologies. *Cahiers du C.I.E.L.: Problèmes de classement des unités lexicales*, 11–48.

Tsitsanoudi-Mallidi, N. (2020). *Journalistic Discourse. From the Metapolitefsi to the "Memoranda" [Ο Δημοσιογραφικός Λόγος. Από τη Μεταπολίτευση έως τα «Μνημόνια»]*. Gutenberg.

Contact information

Nikos Mathioudakis
Research Centre for Scientific Terms and Neologisms, Academy of Athens
nikosmathious@gmail.com

Alexandra Fiotaki
University of Athens
alexandra.fiotaki@gmail.com

Ondřej Matuška, Miloš Jakubíček & Vojtěch Kovář

MANAGEMENT OF LARGE LEXICOGRAPHIC PROJECTS WITH LEXONOMY

Keywords Lexicographic workflows; dictionary drafting; large-scale dictionary projects; post-editing; human validation; collaborative editing

Modern lexicographic workflows increasingly combine automatic dictionary drafting from large corpora with extensive human post-editing and validation. Advances in corpus linguistics and natural language processing make it possible to generate comprehensive dictionary drafts automatically, but the quality of the final product still depends on systematic human editorial work. As a result, contemporary dictionary projects are characterized by multiple iterative cycles in which editors assess, correct, enrich, and approve automatically generated entries. While this approach significantly accelerates dictionary production, it also introduces new challenges in organizing editorial work, maintaining consistency, and supervising progress across large editorial teams.

In this paper we present Lexonomy, a system designed to support the management of large lexicographic projects following this draft-and-post-edit paradigm. The system is tailored for dictionaries whose editors working in parallel. It supports the division of editorial labour into clearly defined tasks where each editor or team focuses on a concrete entry element according to their expertise and lexicographic experience. Lexonomy enables editors to focus on their lexicographic decisions while ensuring that individual contributions fit coherently into the overall editorial plan.

A key objective of the system is to provide the editor-in-chief with a clear and continuously updated overview of the dictionary creation process. By integrating support for collaborative editing with tools for project supervision, Lexonomy facilitates efficient management of large-scale dictionary projects and helps ensure that automatically generated drafts are systematically transformed into high-quality, human-validated lexicographic resources.

Contact information

Ondřej Matuška
Lexical Computing
ondrej.matuska@sketchengine.eu

Miloš Jakubíček
Lexical Computing
milos.jakubicek@sketchengine.eu

Vojtěch Kovář
Lexical Computing
vojtech.kovar@sketchengine.eu

Cordula Meißner, Anna-Lena Randermann & Janina Deilke

INDEXICALITY IN THE VOCABULARY OF GERMAN COMMUNICATION VERBS

Examining Language Use and Speaker Knowledge as a Basis for Lexicographical Conception

Keywords Indexicality; vocabulary of German communication verbs; language use vs. knowledge of language; lexicological description; onomasiological structure

Speakers may need to consciously assess which expressions are appropriate for a communicative situation. One such dimension is indexicality: linguistic expressions may evoke the situational contexts in which they are typically employed. In a usage-based understanding, these indexical potentials emerge from recurrent co-occurrence between linguistic forms and situational features and become part of speaker knowledge through entrenchment and schematization (Schmid, 2020). Corpus analysis is methodologically important because natural language frequency patterns can be treated as a model of the input from which speakers may abstract linguistic knowledge (Stefanowitsch & Flach, 2017; Ellis, 2017). However, existing lexicographical resources do not systematically represent usage-based indexicality, and studies that triangulate corpus-derived patterns with speaker knowledge at the vocabulary level are missing. The project introduced in this paper aims to describe systematically and model usage-based indexicality in German communication verbs (CVs) (Harras et al., 2004), drawing on data from both language use and speaker knowledge. Its research approach can provide a basis for conceptualizing the representation of indexicality in lexicographical resources. At the current stage, the paper focuses on the methodological design of the project (first results will be available when the conference takes place).

The project builds on a preliminary corpus study of spoken German (Meißner, 2025), in which CVs were investigated in the FOLK corpus across private, institutional, and public interaction domains (Kaiser, 2018). Analyses of variance and distinctive collexeme analyses (Gries & Stefanowitsch, 2004) were used to identify patterns of over- and under-representation of CV lexemes across domains (indexical potential types (IPTs)). The project asks to what extent such IPTs correspond to speakers' associations, both with respect to fields of CVs with similar indexical profiles and in terms of granularity, i.e., whether indexical meaning attaches not only to lexemes but also to specific grammatical forms. To this end, three forms are selected: first-person singular present, third-person singular present, and impersonal passive, representing speaker-centered, reporting, and agent-detached perspectives.

To answer the question, corpus-based evidence is triangulated with experimental data on speaker knowledge. The corpus component is based on twelve subcorpora, each containing 100,000 full-verb occurrences, designed to represent contemporary German across private, institutional, and public communication and across spoken, written, computer-mediated, and selected special types of communication. Existing corpus resources are used where possible and are supplemented with new web data where necessary. After preprocessing, analysis of variance and distinctive collexeme analysis are used to derive IPTs for lexemes and grammatical forms.

The speaker-knowledge component mirrors this corpus-based perspective through online questionnaires. For lexemes, 605 CVs from Harras et al. (2004) are distributed over ten item lists and tested in five questionnaire designs: one general association-strength rating,

three domain-specific ratings for private, institutional, and public communication, and one free-association task. The rating tasks use labelled seven-point scales; the free-association task asks which communicative situation first comes to mind. This combines scaled judgment data with less pre-structured elicitation of prototypical contexts. A form-level study mirrors these designs for a 20% subset of 121 CVs for the grammatical forms. The rating data are analyzed with ordinal mixed models, with rating type, form, and their interaction as fixed effects and random intercepts for participants and items.

By comparing corpus-derived IPTs with rating and association profiles, the study identifies convergences, divergences, and complementary patterns between language use and speaker knowledge. It refines a usage-based account of indexicality by examining how frequency-based distributional patterns relate to metalinguistic associations. Thereby, the study provides an empirical basis for conceptualizing a representation of situational-indexical information in lexicographical resources for German CVs. More generally, the project outlines a transferable method for modelling how expressions are associated with communicative situations. A lexicographical representation of such associations can support users in choosing forms appropriate to specific communicative settings.

References

Ellis, N. (2017). Cognition, Corpora, and Computing: Triangulating Research in Usage-Based Language Learning. *Language Learning*, 67(S1), 40–65. https://doi.org/10.1111/lang.12215

Gries, S. T [Stefan Th.], & Stefanowitsch, A. (2004). Extending collostructional analysis. *International Journal of Corpus Linguistics*, 9(1), 97–129. https://doi.org/10.1075/ijcl.9.1.06gri

Harras, G., Winkler, E., Erb, S., & Proost, K. (2004). *Handbuch deutscher Kommunikationsverben: Teil 1: Wörterbuch*. Schriften des Instituts für Deutsche Sprache: 10.1. Walter de Gruyter. https://doi.org/10.1515/9783110201659

Kaiser, J. (2018). Zur Stratifikation des FOLK-Korpus: Konzeption und Strategien. *Gesprächsforschung*, 19, 515–552. https://ids-pub.bsz-bw.de/frontdoor/index/index/year/2019/docId/8668

Meißner, C. (2025). Gebrauchsbasierte Indizierungspotenziale und diasituative Variation: Die Nutzung von Kommunikationsverben im Spektrum des gesprochenen Deutsch. In S. Stumpf, M.-L. Merten, S. Kabatnik, & S. Zollner (Eds.), *Studien zur Pragmatik: Vol. 8. Variationspragmatik: Regionale Vielfalt und situative Unterschiede im Sprachgebrauch* (pp. 257–297). Narr Francke Attempto. https://doi.org/10.24053/9783381135127

Schmid, H.-J. (2020). *The dynamics of the linguistic system: Usage, conventionalization, and entrenchment*. Oxford University Press. https://doi.org/10.1093/oso/9780198814771.001.0001

Stefanowitsch, A., & Flach, S. (2017). The corpus-based perspective on entrenchment. In H. J. Schmid (Ed.), *Language and the human lifespan series. Entrenchment and the psychology of language learning: How we reorganize and adapt linguistic knowledge* (pp. 101–128). De Gruyter Mouton; American Psychological Association.

Contact information

Cordula Meißner
Universität Innsbruck
cordula.meißner@uibk.ac.at

Anna-Lena Randermann
Universität Innsbruck
anna-lena.randermann@uibk.ac.at

Janina Deilke
Universität Innsbruck
janina.deilke@uibk.ac.at

Olga Menagarishvili

DICTIONARIES OF ENGINEERING IN THE AGE OF AI

Keywords Dictionaries of engineering; GenAI definitions; technical communication; cultural studies

As cultural objects, dictionaries perform several types of cultural work. For example, they have the ability to influence knowledge legitimation: they exercise power as they assign value to knowledge by dictating what knowledge is valuable and trustworthy (the knowledge found in dictionaries) and what is not (the knowledge omitted by dictionaries) (Menagarishvili, 2020). Dictionaries also create what Anderson (2006, p. 6) calls "imagined communities" that can be defined as communities consisting of people who "will never know most of their fellow-members, meet them, or even hear of them, yet in the minds of each lives the image of their communion." Such communities are present in our lives even though most of the time we are unaware of them, and this invisibility, perhaps, makes imagined communities even more powerful. Finally, each dictionary is created within a certain cultural context that affects that dictionary's content, design, and usage.

As artifacts of technical communication, dictionaries are a genre with many subgenres that can be used to communicate about technology. Each dictionary has an audience, a certain purpose, and a structure and is used in a certain way, which are all aspects of documents discussed by scientific and technical communication (Menagarishvili, 2023). Moreover, definitions are a typical technical communication genre that is taught in every technical communication class. Further, many dictionaries combine text and visuals, which is another technical communication topic taught and studied by technical communication instructors/researchers.

In this poster presentation, I will discuss dictionaries of civil and environmental engineering as cultural objects and as artifacts of technical communication. More specifically, I will compare the following dictionaries describing civil and environmental engineering discourse: *The McGraw-Hill Dictionary of Scientific and Technical Terms*; McGraw-Hill's *Dictionary of Engineering*; *A Dictionary of Civil and Environmental Engineering*; *A Dictionary of Construction, Surveying, and Civil Engineering*; and *Environmental Engineering Dictionary and Directory*. The field of civil and environmental engineering was chosen for analysis because, as an Engineering Communication Program Director in a School of Civil and Environmental Engineering, I observed that dictionaries describing this vibrant field have not been discussed in the literature. These specific dictionaries were selected for analysis because they all describe civil and/or environmental engineering terminology but differ in their level of field specificity. They are listed above in order of increasing field specificity: technical terms–engineering terms–civil and environmental engineering terms–civil engineering-only terms / environmental engineering-only terms. This part of the study will include dictionary macrostructure and microstructure analysis.

In the second part of the study, I will explore what happens from the cultural studies and technical communication perspectives when civil and environmental engineering terms make the jump from traditional dictionaries to GenAI. Based on consultations with six civil and environmental engineering experts at a top-ranked technological university, I will select five widely used civil and environmental engineering terms from each of the six subfields of civil and environmental engineering (30 terms total). The six subfields include construction and infrastructure engineering; geosystems engineering; structural engineering, mechanics, and materials; transportation systems engineering; water resources engineering; and environmental engineering. I will then conduct a cultural and lexicographic analysis (1) of dictionary articles from

the dictionaries listed above and (2) of definitions provided by five popular GenAI tools likely to be used by civil and environmental engineering students at a prestigious technological university in the US in the summer/fall 2026.

This poster presentation might be useful for lexicographers, instructors using dictionaries and/or GenAI in their classes, and the general public interested in dictionaries and/or GenAI.

References

Anderson, B. (2006). *Imagined Communities: Reflections on the Origin and Spread of Nationalism.* Verso.

Dictionary of engineering (2003). The McGraw-Hill Companies.

Friebel, H. (2020). *A dictionary of civil and environmental engineering.* Golden Ratio Publishing.

Gorse, Ch., Johnston, D., & Pritchard, M. (2020). *A dictionary of construction, surveying, and civil engineering.* Oxford University Press.

McGraw-Hill dictionary of scientific and technical terms. (2003). New York: McGraw-Hill Book Company.

Menagarishvili, O. (2020). Dictionaries of science and technology and issues of power. *English for Specific Purposes (ESP) and the Underlying Dynamics of Power, Empowerment and Disempowerment*, special issue of *ILCEA (Institute for Languages and Cultures of Europe, America, Africa, Asia and Australia) Online*, 40. https://doi.org/10.4000/ilcea.10614

Menagarishvili, O. (2023, May 31-June 3). What do cultural studies and scientific and technical communication have to do with dictionaries? [Conference presentation]. 24th Biennial Dictionary Society of North America Conference, University of Colorado Boulder, Boulder, CO, United States.

Pankratz, Th. M. (2001). *Environmental engineering dictionary and directory.* CRC Press.

Contact information

Olga Menagarishvili
Georgia Institute of Technology
olga.menagarishvili@ce.gatech.edu

Tomasz Michta & Tomasz Czyż

CHATGPT OR A DICTIONARY?

Resolving Polysemy in L2 Reception

Keywords ChatGPT; large language models; polysemy; user study

Recent empirical studies suggest that ChatGPT can assist users with a range of receptive and productive language tasks (Rees & Lew, 2023; Ptasznik, Wolfer & Lew, 2024). However, users consult dictionaries in many situations not covered in previous studies comparing ChatGPT with dictionaries, and ChatGPT's usefulness is likely to vary depending on the task. This is plausible because ChatGPT does not appear to perform equally well when asked to produce different types of dictionary information: it has been found to produce definitions that are "practically indistinguishable" from those found in the COBUILD dictionary (Lew, 2023, p. 8), while ChatGPT-generated examples tend to be repetitive, unimaginative and unnatural (Jakubíček & Rundell, 2023; Lew, 2023). The present study focuses on a situation often faced by language learners but still underexplored in empirical comparisons of ChatGPT and dictionaries: resolving polysemy in L2 reception.

Polysemous L2 words pose a considerable challenge for language learners because they can be deceptively familiar: learners may assume they understand a word because they know one of its senses, although the context requires another. If they do consult a dictionary, they still have to determine which sense in the entry matches the context. As earlier studies have shown, many learners struggle to do so (Dziemianko, 2019; Kamiński, 2025).

Learners may now also use an LLM to help them understand which sense of a polysemous word is intended in context. LLMs may be useful here, but they may also present learners with false polysemy by proposing senses that are not clearly distinct from one another (Jakubíček & Rundell, 2023). This matters because false polysemy has been shown to make sense selection more time-consuming and to lower users' confidence in the sense they selected (Michta & Frankenberg-Garcia, 2025).

In this study, we examine how successfully learners identify the contextually appropriate senses of polysemous L2 items. To this end, we compare the effectiveness of ChatGPT with that of a bilingual dictionary in an equivalent-selection task and with that of a monolingual learner's dictionary in a definition-selection task. In addition to accuracy, we investigate learners' confidence in their choices, time-on-task, and order effects.

The participants were 131 L1 Polish secondary-school students whose proficiency in English ranged from B1 to B2. Fifteen polysemous English nouns were selected as target items. Each item appeared in a single sentence taken from Sketch Engine for Language Learning, published books, or dictionaries other than those used by participants in the study. Some sentences were minimally edited to remove non-essential information. The task was administered via a worksheet in which sentences were presented in random order.

The study consisted of a pretest and a main test. In the pretest, all participants saw 15 sentences and attempted to translate or paraphrase each target item without tool support. In the main test, participants were assigned to one of four groups and used their phones to complete one of two tasks. In the equivalent-selection task, Group A used Diki.pl, an online English-Polish dictionary, while Group B used ChatGPT. Participants selected a Polish equivalent of the target word and rated their confidence in selecting an appropriate equivalent on a 1–5 Likert scale. In the definition-selection task, Group C used the online Oxford Advanced Learner's Dictionary,

while Group D used ChatGPT. Participants indicated their selected definition by copying three words from it and rated their confidence in their choice on a 1–5 Likert scale. They were then asked to provide a Polish equivalent of the English target word. Total time-on-task was recorded for the main test.

The data were analysed separately for each of the two tasks using logistic mixed-effects models for accuracy and ordinal mixed-effects models for confidence. Preliminary results suggest that, in the equivalent-selection task, Diki.pl users outperformed ChatGPT users and reported higher confidence. Order effects were also observed in this task. Full analyses will be presented at the conference.

References

Dziemianko, A. (2019). Homogeneous or Heterogeneous? Insights into Signposts in Learners' Dictionaries. *International Journal of Lexicography*, 32(4), 432–457. https://doi.org/10.1093/ijl/ecz011

Jakubíček, 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. Jakubíček (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.

Kamiński, M. (2025). L1 Signposts are Better for Sense Selection, Translation Accuracy, and Consultation Speed than L2 Signposts. *International Journal of Lexicography*, 38(4), 321–341. https://doi.org/10.1093/ijl/ecaf018

Lew, R. (2023). ChatGPT as a COBUILD lexicographer. *Humanities and Social Sciences Communications*, 10(704), 1–9. https://doi.org/10.1057/s41599-023-02119-6

Michta, T. & Frankenberg-Garcia, A. (2025). Learners' reactions to false polysemy. In I. Kosem, M. Jakubíček, M. Medveď, K. Zgaga, Š. Arhar Holdt, T. Munda & A. Salgado (Eds.), *eLex 2025: Book of abstracts* (pp. 60–61). Ljubljana University Press.

Ptasznik, B., Wolfer, S. & Lew, R. (2024). A Learners' Dictionary Versus ChatGPT in Receptive and Productive Lexical Tasks. *International Journal of Lexicography*, 37(3), 322–336. https://doi.org/10.1093/ijl/ecae011

Rees, G. P. & Lew, R. (2023). 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

Contact information

Tomasz Michta
University of Bialystok
t.michta@uwb.edu.pl

Tomasz Czyż
University of Bialystok
c86861@student.uwb.edu.pl

Tomasz Michta & Katarzyna Mroczyńska

SPOILT FOR CHOICE?

ChatGPT versus a Bilingual Dictionary in Selecting L2 Equivalents for Polysemous L1 Words

Keywords Bilingual dictionaries; ChatGPT; large language models; equivalence; polysemy; user study

A common challenge for language learners is finding appropriate L2 equivalents for L1 words. The difficulty is compounded when the L1 word is polysemous, because different senses typically call for different equivalents. When selecting an L2 equivalent for a polysemous L1 word, learners therefore face a twofold task: identifying the intended sense of the L1 word and choosing an L2 equivalent that fits the context.

Traditionally, learners have dealt with this problem by turning to bilingual dictionaries. High-quality bilingual dictionaries may provide glosses, labels, collocations, examples, and other information that helps users select the most appropriate equivalent from among those listed and then use it in context. Yet, as previous studies have shown, the mere availability of information does not guarantee that users will find it or use it effectively (Lew & Tokarek, 2010; Dziemianko, 2012).

With the arrival of ChatGPT and other large language models (LLMs), learners may now turn to AI tools instead of consulting dictionaries. A recent study by Ptasznik and Lew (2025) indicates that at least some learners already do so, although their preferences appear to be task-dependent. Learners' willingness to use such tools is perhaps unsurprising, given evidence from user studies showing that LLMs can match and sometimes outperform dictionaries in receptive and productive tasks (Rees & Lew, 2023; Ptasznik, Wolfer & Lew, 2024). Yet expert evaluations of entries generated by LLMs have been mixed. An often-noted problem is false polysemy, which occurs when "the system enumerates multiple senses, with different definitions, in cases where there is really only one" (Jakubíček & Rundell, 2023, p. 525). A recent user study found that false polysemy can have a detrimental effect on dictionary users, increasing the time needed to select a definition and lowering users' confidence in the definition chosen (Michta & Frankenberg-Garcia, 2025).

To the best of our knowledge, no study to date has specifically examined learners' performance in selecting L2 equivalents for polysemous L1 words when consulting a bilingual dictionary versus an LLM, a task that is common yet under-researched. We investigate learners' performance using the following measures: response accuracy, response time, and delayed recall of the target equivalents one week later. In addition, we examine learners' confidence in the equivalents they selected and changes in their tool preferences after the task.

A total of 105 L1 Polish speakers, all enrolled in an English philology programme, took part in the study. Before completing the task, participants indicated which of the two tools, ChatGPT or the online Polish-English PONS dictionary, they would prefer to use for the task and rated how useful they expected each tool to be. They were then randomly assigned to one of the two conditions. Participants completed an online task involving 20 sentences presented in random order. Each item consisted of an English sentence with a gap and a Polish polysemous word in parentheses. Participants first indicated whether they were able to supply an English equivalent without consulting the assigned tool, and if so, provided it. Next, they consulted the assigned tool and filled in the gap using the information obtained during consultation; response time was recorded. After each response, they rated how confident they were that they had

selected an appropriate equivalent. One week later, 70 participants completed a post-test in which the same Polish words and senses were presented to them in new contexts. For each item, they provided an English equivalent and rated their confidence. Participants also rated the overall helpfulness of the assigned tool after both the main task and the post-test.

The results suggest that the dictionary group provided more accurate responses, required less time to supply an equivalent, and reported higher confidence in their choices. After the task, participants more often selected the dictionary than ChatGPT as the tool they would choose for a similar task in the future. Full analyses will be presented at the conference.

References

Dziemianko, A. (2012). Why One and Two Do Not Make Three: Dictionary Form Revisited, *Lexikos,* 22, 195–216. https://doi.org/10.5788/22-1-1003

Jakubíček, 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. Jakubíček (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. & Tokarek, P. (2010). Entry menus in bilingual electronic dictionaries. In S. Granger & M. Paquot (Eds.), *eLexicography in the 21st century: New challenges, new applications* (pp. 193–202). Cahiers du CENTAL.

Michta, T. & Frankenberg-Garcia, A. (2025). Learners' reactions to false polysemy. In I. Kosem, M. Jakubíček, M. Medveď, K. Zgaga, Š. Arhar Holdt, T. Munda & A. Salgado (Eds.), *eLex 2025: Book of abstracts* (pp. 60–61). Ljubljana University Press.

Ptasznik, B. & Lew, R. (2025). Dictionaries versus AI Tools through the Eyes of English Majors. *International Journal of Lexicography*, 38(2), 140–158. https://osf.io/vk5cs/files/9wrf6

Ptasznik, B., Wolfer, S. & Lew, R. (2024). A Learners' Dictionary Versus ChatGPT in Receptive and Productive Lexical Tasks. *International Journal of Lexicography*, 37(3), 322–336. https://doi.org/10.1093/ijl/ecae011

Rees, G. P. & Lew, R. (2023). 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

Contact information

Tomasz Michta
University of Bialystok
t.michta@uwb.edu.pl

Katarzyna Mroczyńska
University of Siedlce
katarzyna.mroczynska@uws.edu.pl

Tomasz Michta, Geraint Paul Rees & Ana Frankenberg-Garcia

CORPUS-BASED VERSUS AI-GENERATED DICTIONARY EXAMPLES

How They Differ and What Learners Prefer

Keywords Artificial intelligence; dictionary examples; user study; user evaluation

A distinguishing feature of English dictionaries for learners is that most entries include example sentences or phrases that have been copied or adapted from corpora. Their function is to reinforce meaning by showing how words have been used in context, prioritising typical grammar patterns and lexical collocations (Fox, 1987; Frankenberg-Garcia, 2014; Frankenberg-Garcia et al., 2021). While early corpus-based lexicographers had to scan concordances line by line to select good dictionary examples (Krishnamurthy, 1987), nowadays tools like Word Sketches (Kilgarriff et al., 2014) and GDEX (Kilgarriff et al., 2008) shorten the time it takes for lexicographers to find suitable corpus examples. Despite these welcome developments, it is now arguably easier and faster to generate dictionary examples using LLMs. However, in early experiments evaluating AI-generated examples, experts found them to be redundant, unimaginative and inauthentic (Lew, 2023; Jakubíček & Rundell, 2023), although improvements could be seen after some prompt fine-tuning (Lew, 2023). In this study, we wanted to find out how language learners rated AI-generated and corpus-based dictionary examples, to come to a better understanding of how they differ, and to consider the practical implications for lexicography.

Over 200 university students pursuing various degrees (language and non-language) participated in an online task involving 15 monosemous lexical items (from to different part-of-speech categories) that were likely to be unknown. For each lexical item, the participants were presented with the definitions from the Reverso English Dictionary (AI-generated) and the Oxford Advanced Learner's Dictionary (corpus-based) followed by one example from each dictionary. As most entries contained more than one example, the ones presented to each student were randomly selected from the totality of examples available. The participants were asked to rate the 15 AI-generated and the 15 corpus-based examples they saw on a 1-5 scale (poor to excellent). The order of lexical items, definitions and examples shown to each student were all randomized. At the end of the task, the participants were asked to explain what made them give examples high or low ratings. They then responded to a few demographic questions and completed a standardized vocabulary test (LexTALE, Lemhöfer & Broersma, 2012).

Prior to the data collection, the authors of this study conducted a data-driven appraisal of all 63 examples in the dataset. By comparing interpretive coding criteria and discussing divergences, we collaboratively negotiated a systematic analytical framework for describing dictionary examples. Using this framework, we noted a number of significant differences between the AI and the corpus-based examples (e.g., in the corpus-based examples, there was significantly more variability in the number of words and syntax used). However, not all differences were found to be significant (e.g., the presence of contextual cues about meaning).

With regard to the student ratings, overall preliminary findings indicate that there was no significant difference between examples in each resource. However, when only complete sentences were taken into consideration, the Oxford examples were rated significantly more positively.

The full results of the study, including triangulation of our analytical framework for describing dictionary examples along with user ratings and student demographics will be presented at the conference. We believe the study will help to shed new light on how corpus-based and AI-generated dictionary examples can be refined in the future.

References

Frankenberg-Garcia, A. (2014). The use of corpus examples for language comprehension and production. *ReCALL*, 26(2), 128–146. https://doi.org/10.1017/S0958344014000093

Frankenberg-Garcia, A., Rees, G. & Lew, R. (2021). Slipping Through the Cracks in e-Lexicography. *International Journal of Lexicography*, 34(2), 206–234. https://doi.org/10.1093/ijl/ecaa022

Jakubíček, 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. Jakubíček (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.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. & Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography*, 1(1), 7–36. https://doi.org/10.1007/s40607-014-0009-9

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M. & Rychlý, P. (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). Institut Universitari de Linguistica Aplicada, Universitat Pompeu Fabra.

Krishnamurthy, R. (1987). The process of compilation. In J. Sinclair (Ed.), *Looking Up: An Account of the COBUILD Project in Lexical Computing* (pp. 62–85). Collins ELT.

Lemhöfer, K. & Broersma, M. (2012). Introducing LexTALE: A quick and valid lexical test for advanced learners of English. *Behavior Research Methods*, 44, 325–343. https://doi.org/10.3758/s13428-011-0146-0

Lew, R. (2023). ChatGPT as a COBUILD lexicographer. *Humanities and Social Sciences Communications*, 10(704), 1–10. https://doi.org/10.1057/s41599-023-02119-6

Contact information

Tomasz Michta
University of Bialystok
t.michta@uwb.edu.pl

Geraint Paul Rees
Universitat Pompeu Fabra
geraintpaul.rees@upf.edu

Ana Frankenberg-Garcia
University of Surrey
a.frankenberg-garcia@surrey.ac.uk

Julie Moore

DIRECTIONS IN LEXICOGRAPHY IN THE GLOBAL SOUTH

Keywords Global South; minority languages; translanguaging; ELT; ESP; EMI; inclusive lexicography; AS Hornby Trust

The AS Hornby Dictionary Research Awards (ASHDRA) have been running since 2019 and to date have received 160 applications from 47 countries. The awards fund dictionary research within English language teaching (ELT) and especially target projects in low-resource contexts. Thus, a large proportion of applications (although not all) are from researchers working in the Global South. This paper reviews three key trends that have emerged across those applications and suggests areas of particular focus in lexicography in these regions, with examples of ASHDRA-funded projects in each area.

Multilingual resources including local languages

The most significant trend to emerge among ASHDRA research proposals has been interest in bi- and multilingual reference resources for minority, local languages. Many countries of the Global South have a multi-layered linguistic patchwork in which people speak a local language at home as their L1, then learn a regional or national language as their L2 and lingua franca. English (effectively their L3) is commonly taught from primary school but typically via the L2. In one such project, Riadi (2021) created a multilingual dictionary for learners in schools in an underdeveloped region of West Borneo, Indonesia. The small-scale resource was based around photographs of the linguistic landscape (adverts, shop signs, etc.) near to the participating schools. The entry for each target word included equivalents in English, Bahasa Indonesia (the national language), Pontianak (a regional language), and Ketapang (a local language and most students' mother tongue). The resulting dictionary provided both a classroom resource enabling students to translate directly between English and their home language, something they had not previously done (or been encouraged to do) and also linked the English they were learning to their environment and lived experiences, something that classroom observation and feedback showed they found incredibly motivating.

Other projects (proposed and ongoing) have focused on communities in rural India who speak a local language at home, then learn English at school through the medium of a regional language, such as Kannada, or indigenous students in Mexico who speak Zapotec, a largely oral local language, as their home language but are schooled in Spanish, then are expected to learn English. In some cases, schoolteachers do not come from the local communities and so do not even share their students' L1. This puts those learners, who may already come from disadvantaged communities, at a further disadvantage. With an increasing interest in 'translanguaging' and the benefits of using learners' L1 as a resource in language teaching, some national Ministries of Education are encouraging the use of L1 in the ELT classroom. Such an approach is hard to implement, however, without relevant resources for both learners and teachers. Around 25% of ASHDRA applications have been to develop multilingual resources for contexts where none yet exist. These projects often come up against the further barrier that local languages may be largely oral and even where written forms exist, there may be low levels of literacy within the community. This has led to creative solutions such as the use of images and audio (see more below).

English for (very) specific purposes

English for Specific Purposes (ESP) is a well-established area of lexicography, but many published resources, such as dictionaries of Medical English or English for Engineering, are quite broad in scope and often geared toward typical students and teaching contexts in countries in the Global North. Despite the massive growth in English as a medium of instruction (EMI) in higher education globally, the differing needs of learners in the Global South remain under-researched (Sahan et al., 2021) and existing ESP resources may not always be relevant or accessible to learners in low-resource, sometimes low-bandwidth, contexts with specific English needs. Another theme among ASHDRA proposals has been the development of very specific resources to target the needs of particular groups of English learners, such as a dictionary of semi-technical vocabulary aimed at Vietnamese medical students (Le, 2022), an app for Omani physics undergraduates (Mathew & McCallum, 2024), or a resource for Indonesian mining engineering students learning to work with the industry standard software in English (Nurfitriah, forthcoming). With limited resources available, these projects are not necessarily full ESP dictionaries but resources with a scope and format targeted at the specific and immediate needs of learners.

Inclusive lexicographic resources

The final theme to emerge is the development of accessible resources for language learners who cannot easily access conventional dictionaries. This group includes learners with visual and hearing impairments. Two research projects due for completion in 2026 target, respectively, visually impaired students in Indonesian primary schools with a dictionary that includes both Braille and QR codes linked to audio (Karolina, forthcoming), and a project using sign language videos aimed at deaf learners in India (Zeshan, forthcoming). This category also includes learners with low levels of literacy. This overlaps with some of the projects under the first theme where local languages may be from a largely oral tradition, but also adult migrant and refugee learners who sometimes have low levels of literacy in their first language and who can more easily access English language resources with visual and/or audio components (such as O'Boyle, 2023, Oakey & Jones, forthcoming).

Trends: focusing on the local and specific

What the most innovative projects in all three areas have in common is that they take as their starting point the needs of learners in a specific context and often work in collaboration with learners and educators to develop a lexicographic resource that directly addresses those needs, working with limited resources in creative ways. While it has to be acknowledged that ASHDRA have a specific focus and the proposals received reflect this, it seems that they may indicate a more general trend in lexicography and the projects they fund may provide models which could be replicated elsewhere.

References

Karolina, V. (forthcoming). *Braille English school dictionary with interactive audio-based AI support for visually impaired primary school students in Indonesia*. Hornby Trust. https://www.hornby-trust.org.uk/a-s-hornby-dictionary-research-awards-ashdra/

Le, C. (2022). *Piloting SemiMed – a mini semantic visualisation dictionary of semitechnical medical terms*. Hornby Trust. https://www.hornby-trust.org.uk/a-s-hornby-dictionary-research-awards-ashdra/

Mathew, P. & McCallum, L. (2024). *A bilingual corpus-based physics dictionary for L1 Arabic undergraduates*. Hornby Trust. https://www.hornby-trust.org.uk/a-s-hornby-dictionary-research-awards-ashdra/

Nurfitriah. (forthcoming). *Enhancing English terminology resources for mining engineering students and professionals through AI-assisted lexicography*. Hornby Trust. https://www.hornby-trust.org.uk/a-s-hornby-dictionary-research-awards-ashdra/

Oakey, D. & Jones, C. (forthcoming). *An open-access learner-led online spoken English lexicographic resource for improving spoken vocabulary use in adult learners of English as an additional language (EAL)*. Hornby Trust. https://www.hornby-trust.org.uk/a-s-hornby-dictionary-research-awards-ashdra/

O'Boyle, A. (2023). *A welcome dictionary: refugee families using dictionary resources to support English language learning*. Hornby Trust. https://www.hornby-trust.org.uk/a-s-hornby-dictionary-research-awards-ashdra/

Riadi, A. (2021). *Incorporating linguistic landscape in a plurilingual dictionary to support learners' translanguaging in a Borneo underdeveloped region: a preliminary pilot study*. Hornby Trust https://www.hornby-trust.org.uk/a-s-hornby-dictionary-research-awards-ashdra/

Sahan, K., Mikolajewska, A., Heath, R., Macaro, E., Searle, M., Aizawa, I., Zhou, S. & Veitch, A. (2021). *Global Mapping of English as a medium of instruction in higher education: 2020 and beyond*. British Council. https://www.teachingenglish.org.uk/publications/case-studies-insights-and-research/global-mapping-english-medium-instruction-higher

Zeshan, U. (forthcoming). *Bimodal-bilingual function word dictionary for English grammar games with deaf sign language users in India*. Hornby Trust. https://www.hornby-trust.org.uk/a-s-hornby-dictionary-research-awards-ashdra/

Contact information

Julie Moore
AS Hornby Trust
julie@juleswords.co.uk

Phoebe Nicholson, Will Rogers & Kate Wild

HISTORICAL LEXICOGRAPHY IN THE AGE OF AI

A Progress Report from the OED

Keywords Oxford English Dictionary; historical lexicography; AI-enabled lexicography

The Oxford English Dictionary (OED) has long been recognized for its rigorous scholarship, yet it is equally notable for its readiness to adopt new technologies. As the volume and availability of linguistic evidence has grown in the digital era, so too have the challenges of efficiently revising a dictionary of this size and scope: in response, the OED has been undertaking a systematic investigation into the potential role of AI in supporting historical lexicography, aspects of which have been reported at previous conferences (Hawkes, Nicholson & Rogers, 2025; Wild, 2024). In this paper we present findings to date on successful and less successful experiments, and describe the OED's approach to scaling up the successes.

Much has been written about the capabilities of generative AI to draft dictionary definitions; in this paper, we report on two experiments that instead explore how LLMs can support the adaptation or updating of existing definitions for different purposes or audiences. The first experiment involved adapting definitions from the OED (a large historical dictionary of English) into a style suitable for the Oxford Dictionary of English (a general-purpose dictionary of current English). This requires removing obsolete or minor senses, making technical or formal language more accessible, simplifying long definitions, and similar transformations. We tested whether prompting an LLM to perform this task, followed by editorial review, could yield consistent and time-saving results; however, the output proved too variable to be implemented at scale.

The second experiment (initially reported in Wild, 2024) focused on modernizing unrevised OED definitions, many of which were written in the 19th century, into a style appropriate for contemporary readers. With over half a million entries and more than 800,000 senses, the OED is one of the largest dictionaries in existence. This presents significant challenges in determining which entries and senses are most in need of updating. We first discuss methods of identifying outdated definitions, including traditional computational methods (such as finding definitions containing words which have declined in frequency in a corpus), and AI-supported methods (including diachronic embeddings and LLM verification). We then present the results of an experiment using an LLM to 'translate' archaic wording into modern English, when supplied with the unrevised OED definition, its associated illustrative quotations, and a prompt with examples. The results showed that AI-modernized definitions can serve as an aid for editors, along with other sources, although editorial input remains essential.

We have also investigated the capacity for AI to assist with retrieving and analyzing quotation evidence. OED's rich historical sense inventory is supported by typical, illustrative examples of a word or meaning, including the earliest use of a given sense and a recent example if the sense is still in current use. Research activity on the OED in this area involves laborious sifting of a large quantity of data from primary source databases, and the manual nature of this activity has limited the efficiency of how we work, and the volume of entries we can update. We evaluate the extent to which AI-assisted word sense disambiguation can support this task, firstly in a project using LLMs to retrieve quotations from a corpus of historical English and

match them to OED senses to create a 'Quotations Finder' tool that aims to support the work of OED editors, and supplement OED quotation paragraphs for external users (initially reported in Hawkes, Nicholson & Rogers, 2025). As part of this project, we explored the performance of 7 different LLM models and their comparative ability to carry out word sense disambiguation, and we will describe the relative merits of the model selected for the production of this tool. We also discuss a similar project for modern English which matches OED senses with examples from large corpora of 21st-century English. Our analysis shows that AI-assisted word sense disambiguation (in both historical and modern texts) has a relatively high accuracy rate for certain types of entry and sense, but that there are limitations and areas where editorial intervention is essential in selecting suitable quotations.

Taken together, these projects demonstrate both the promise and the limits of applying AI to historical lexicography. While current models cannot replace the nuanced judgement required for OED revision, they can meaningfully augment editorial work by accelerating manual tasks, and improve the discoverability of evidence. Our findings suggest that, when deployed selectively and with rigorous oversight, AI can help the OED scale its research and modernize its content, while preserving the scholarly standards that define the dictionary.

References

Hawkes, E., Nicholson, P. & Rogers W. (2025). Making Sense of the Past: AI-Assisted Historical Word Sense Disambiguation and the OED. In Kosem, I., Jakubíček, M., Medveď, M., Zgaga, K., Arhar Holdt, Š., Munda, T. & A. Salgado (Eds.), *Electronic lexicography in the 21st century (eLex 2025): Intelligent lexicography. Proceedings of the eLex 2025 conference* (pp. 34-35). Lexical Computing CZ s.r.o.

Wild, K. (2024). Using Large Language Models for a Large Historical Dictionary: challenges and opportunities for the Oxford English Dictionary. In Despot, K. Š., Ostroški Anić, A., & I. Brač (Eds.), *Lexicography and Semantics: Book of Abstracts of the XXI EURALEX International Congress* (pp. 225-227). Institut za hrvatski jezik.

Contact information

Phoebe Nicholson
Oxford University Press
phoebe.nicholson3@oup.com

Will Rogers
Oxford University Press
will.rogers@oup.com

Kate Wild
Oxford University Press
kate.wild@oup.com

Maciej Ogrodniczuk & Dariusz Czerski

CAN LLMS AUTOMATE THE NEOLOGISM LIFE CYCLE?

Evaluating GenAI in Lexicographical Description of New Polish Words

Keywords Large language models; neologisms; e-lexicography

The dynamic development of the modern Polish language, driven by technological and social shifts, poses a significant challenge to traditional lexicography. Traditional methods of documenting neologisms, relying on manual curation and limited dictionary analysis, are increasingly inefficient against the pace of lexical innovation. At the same time new tools such as NeoN (Tomaszewska et al., 2025), built within the ENEOLI COST Action (Tallarico et al., 2025), try to align with modern e-lexicography by utilizing large language models (LLMs) to automate the description of new linguistic units. We evaluate NeoN to see whether it can be reasonably used to generate definitions of new lexical items by carrying out systematic tests of its output with lexicographers and average language users.

The NeoN system implements a multi-layered architecture that combines classic corpus-driven methods with advanced Generative AI (GenAI) which is used to generate definitions of lexical items acquired by continuously monitoring RSS feeds and is subsequently filtered using various corpus-, lexicon- and frequency-based methods. In the current research, we focus on the systematic evaluation of the proposed generation model by validating definitions for a curated set of approximately 50 neologisms collected over a full year (in a 'neologism of the week' setup). This study involves a comprehensive human evaluation phase where annotators judge the quality of definitions produced under various experimental conditions. We test different LLMs and varying levels of provided context to determine their impact on accuracy of generated definitions. Furthermore, we intend to conduct a comparative analysis by providing the same context to human participants, asking them to write definitions, and then evaluating their output against AI-generated results to identify gaps in nuances and lexicographical precision.

While the NeoN system demonstrates the efficiency of GenAI in drafting definitions, this research will also further explore the "Human-in-the-Loop" (HITL) necessity for maintaining lexicographical integrity. We will also analyze the specific linguistic categories where LLMs struggle, such as subtle pragmatic shifts, cultural connotations, and stylistic labeling (e.g., irony or slang). By identifying these "nuance gaps" during our comparative analysis, we intend to propose a refined workflow where AI handles initial structural detection and drafting, while human experts focus on fine-tuning semantic boundaries. This approach ensures that the resulting lexicographic resources remain authoritative while benefiting from the speed of automated processing.

The integration of GenAI within NeoN demonstrates that LLMs can serve as valuable autonomous assistants in the lexicographer's workflow. The positive results from evaluating its effectiveness suggest that this approach could represent a new direction for neology research, where rigorous linguistic analysis is supported by the flexibility of GenAI and subsequently fine-tuned by human experts.

References

ENEOLI COST Action. Retrieved June 18, 2026, from https://eneoli.eu/

Tallarico, G., Costa, R., Salgado, A., Freixa, J., Heylen, K., Humbley, J., Kernerman, I., Klosa-Kückelhaus, A., Ostroški Anić, A., Vezzani, F., Balnat, V., Božović, P., Holdt, Š. A. & Woźniak-Żak, W. (2025). Lexicography and Neology: Building Bridges through a Multilingual Approach with ENEOLI. *Lexikos*, 35(1), 294–317. https://doi.org/10.5788/35-1-2039

Tomaszewska, A., Czerski, D., Żuk, B., & Ogrodniczuk, M. (2025). NeoN: A Tool for Automated Detection, Linguistic and LLM-Driven Analysis of Neologisms in Polish. In M. H. Lees, W. Cai, S. A. Cheong, Y. Su, D. Abramson, J. J. Dongarra & P. M. A. Sloot (Eds.), *Computational Science – ICCS 2025* (pp. 318–326), Cham, Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-97635-3_38

Contact information

Maciej Ogrodniczuk
Institute of Computer Science, Polish Academy of Sciences
maciej.ogrodniczuk@ipipan.waw.pl

Dariusz Czerski
Institute of Computer Science, Polish Academy of Sciences
dariusz.czerski@ipipan.waw.pl

Iryna Ostapova, Mykyta Yablochkov,
Alona Dorozhynska & Iuliia Verbynenko

DEVELOPMENT OF A WEB APPLICATION FOR A MULTILINGUAL TERMINOLOGICAL DICTIONARY

Keywords Computer lexicography; lexicographic database; lexicographic system; parsing; specialised dictionary; parsing

Most national terminological dictionaries in Ukraine are not accessible for automatic data processing, which makes it impossible to use them in digital information systems. The objective of the research is to develop a technology for converting specialized dictionary text into a website with a developed user interface.

The object of the study was "Dictionary of Ukrainian biological terminology" (7,342 entries and about 26,000 terms in Ukrainian, Russian and English) (Grodzinskiy, 2012), that contains definitions, terms polysemy, synonymy, stresses for Slavic languages, and grammatical information.

Since the dictionary text was available in digital publishing format (PDF), no prior digitization was required. Our approach is to step-by-step transform the linear text of a dictionary into a website. The basic steps are as follows:

1. Dictionary text normalization: restoration of the text line that represents the dictionary entry, stress marking, font markers fixation, correction of inevitable publishing errors in the dictionary entry structure, etc. This was the most time-consuming step, and it required manual processing. The text was converted into .doc format. MS Word text processor was used for processing, the result was text in .txt format, in which HTML tags were used to mark substrings, presented in bold and italic.

2. Designing a dictionary lexicographic system model (Shyrokov, 2021; Kupriianov, 2020). This model serves as a basis for building a parsing algorithm, designing a database schema and interface elements. The model was designed based on an analysis of the printed version of dictionary entries markup. Lexicographic systems model methodology allows us to identify all structural elements that can be identified automatically, and to establish connections between them. Each dictionary entry is assigned one universal structure, i.e. any dictionary entry is considered as a derivative of one "template" entry.

3. Construction of an XML schema based on the conceptual lexicographic model.

4. Automatic conversion of dictionary text (.txt format) into an XML document, allowing to explicate all defined structural elements and the connections between them. To automatically mark the dictionary text with XML tags, a program was developed that highlights the elements of the dictionary entry structure. We consider an XML document as a stand-alone product that effectively represents lexicographic data forfurther use for various purposes.

5. Lexicographic database creation. NoSQL (document-oriented databases) was chosen for this (Belkadi, 2021). In the case of relational databases, data is stored as a set of multiple tables and links between them. Working with individual tables as a single object requires a powerful software infrastructure. Moreover, the evolutionary potential of such a digital

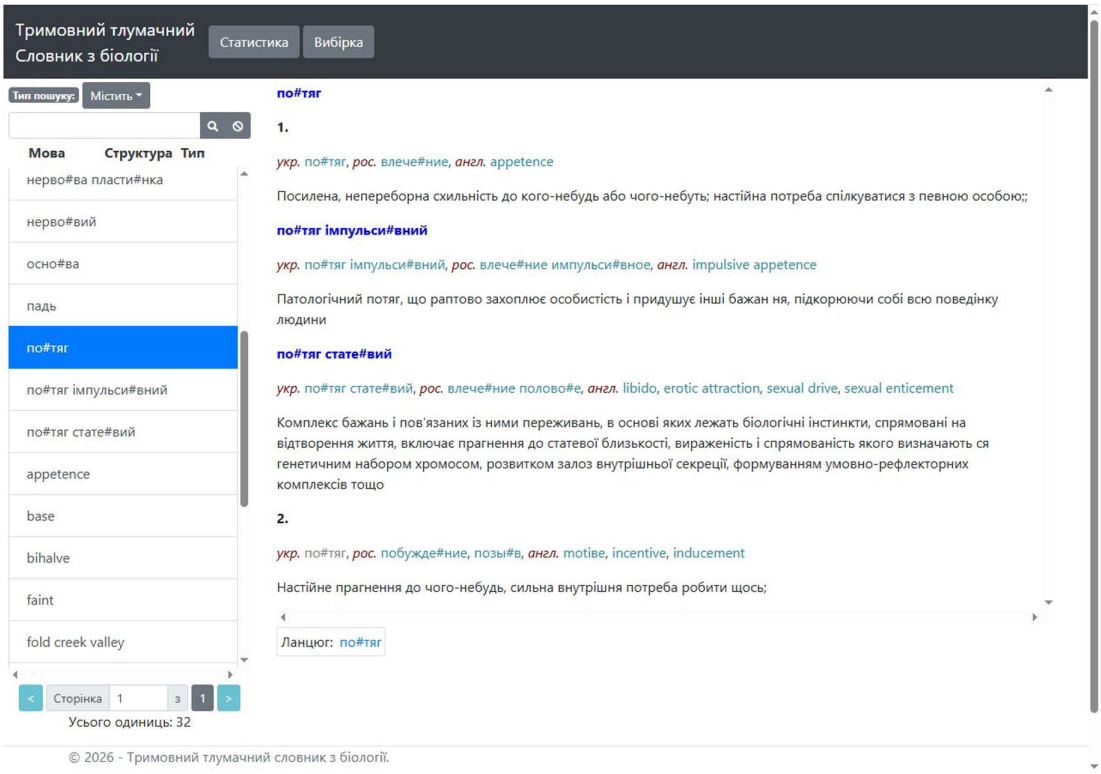


FIGURE 1: Dictionary main window

object is limited by the opacity of the database. Since dictionary entries are the basic elements of a lexicographic system with a strictly defined structure, it is logical to represent them as classes in object-oriented programming languages with subsequent processing, editing and storage in explicit form. The main advantage of NoSQL databases for our project is their ability to store explicitly lexicographic objects without changing their internal structure, which opens direct access to each element of the lexicographic object and significantly simplifies the possibility of editing and modifying (extending) it.

6. Converting XML file to database. This was performed automatically.

7. Designing of interface schemes and creation of a website (currently in progress).

References

Belkadi, F. Zahra & Esbai, R. (2021). A Model-Driven Engineering: From Relational Database to Document-oriented Database in Big Data Context. In *Proceedings of the 16th International Conference on Software Technologies, ICSOFT21*, pp. 653–659.

Grodzinskiy, D. (Ed.) (2012). *Dictionary of Ukrainian biological terminology*. KMM.

Kupriianov, Y., Ostapova, I. & Yablochkov, M. (2020): Design of the User's Interface of Virtual Lexicographic Laboratory for Explanatory Dictionary of the Spanish Language. In S. Lytvyn (Ed.), *Proceedings of the 4th International Conference on Computational Linguistics and Intelligent Systems* (*Volume I*, pp. 121–139). Lviv Polytechnic National University, University of Paris 13, University of Lille, National Technical University "Kharkiv Polytechnical University".

Shyrokov, V. (2021). *Language. Information. System*. Akademperiodyka.

Contact information

Iryna Ostapova
Ukrainian Lingua-Information Foundation of National Academy of Sciences of Ukraine
irinaostapova@gmail.com

Mykyta Yablochkov
Ukrainian Lingua-Information Foundation of National Academy of Sciences of Ukraine
gezartos@gmail.com

Alona Dorozhynska
Ukrainian Lingua-Information Foundation of National Academy of Sciences of Ukraine
alonochkatkachyk@gmail.com

Iuliia Verbynenko
Ukrainian Lingua-Information Foundation of National Academy of Sciences of Ukraine
Yulia_verbynenko@yahoo.com

Ana Ostroški Anić, Federica Vezzani,
Kris Heylen & Ilan Kernerman

FROM DETECTION TO STANDARDIZATION

Practices and Criteria for Terminological Neology

Keywords Lexical innovation; neology; neologisms; terminological neologisms

In terminology planning and terminology work more broadly, the creation and establishment of new terms within specialized domains may appear to be a well-governed practice, guided by widely accepted terminology standards and established procedures across many languages. However, rapid technological development and globalized communication increasingly challenge this model. The continuous creation of English terms, particularly in fast-moving domains, leads to their swift diffusion into other languages, often without adaptation. When adaptation does occur, it frequently follows English morphological patterns and preserves the original, often metaphorical motivation.

These developments raise important questions regarding the identification of emerging terms in texts and other communicative settings, e.g., through corpus-based and AI-assisted term extraction. Criteria for identifying terminological neologisms therefore differ substantially from those applied in lexicography. Whereas lexicographic practice typically relies on frequency as a key indicator distinguishing neologism from individual lexical creativity, in terminology the emergence of a new concept within a specialized field justifies in itself the documentation of a new term, even if its textual frequency remains limited. Consequently, the detection of new terms must combine textual evidence with conceptual and domain-specific validation.

This paper examines current European guidelines and institutional practices concerning neology in terminology, as reflected in relevant research and in the work of terminology institutions. Drawing on data from the ENEOLI survey (Kallas et al., 2025), literature describing national approaches (e.g., Vaus & Janson, 2025; Zanola & Resi, 2025), the recommendations developed within the framework of ISO/TC 37/SC 3, as well as on selected national and institutional materials related to terminology planning (e.g., Prys & Jones, 2007), the study analyses how emerging concepts and their term variants are identified, evaluated, and formalized. Two main research questions are addressed: which criteria are used to determine if a candidate neoterm should be documented in a terminological resource, and what methodological guidelines can be proposed for the detection, description, validation and eventually standardization of neoterms?

The analysis of the responses collected in the ENEOLI survey shows that terminology institutions have not yet developed dedicated policies or formal workflows specifically addressing terminological or lexical innovation. Despite the lack of formalized procedures, several workflows used for the identification and description of new terms have been described, which largely implement the already established principles of terminology work as suggested in ISO standards (ISO 704, 2022). The most frequently mentioned criteria for inclusion of terminological neologisms in terminological resources included domain specificity, conceptual relevance, the frequency of use, corpora attestations, communicative necessity, and the ability to fill lexical or semantic gaps.

In the ISO TC/37 standards, the inclusion of neoterms in terminological resources is also not described by an explicit set of inclusion criteria. Instead, the relevant criteria are distributed

across standards on terminology principles, concept analysis, term formation, harmonization, structured terminological representation, and resource modelling. In a similar vein, the majority of publicly available documents on terminology planning and terminology work that are published by different terminology institutions in Europe largely implement the criteria from ISO standards.

On the basis of the materials collected and analysed, methodological guidelines for the systematic detection and description of terminological neologisms are proposed. The management of terminological neology is understood as an iterative workflow consisting of detection, description, validation, harmonization, dissemination, and revision. These phases form a cycle rather than a one-directional sequence, and they include the application of contemporary corpus-based terminology work alongside generative AI systems, as well as concept analysis and expert validation as the foundations of terminology work.

References

ISO 704:2022 (2022). *Terminology Work—Principles and Methods*. International Organization for Standardization.

Kallas, J., Koppel, K., Heylen, K., Kernerman, I., Ostroški Anić, A., Vezzani, F. & Arhar Holdt, Š. (2025). Neology in Practice: Lexicographic and Terminological Approaches to Lexical Innovation. In I. Kosem, M. Jakubíček, M. Medveď, K. Zgaga, Š. Arhar Holdt, T. Munda & A. Salgado (Eds.), *Electronic lexicography in the 21st century (eLex 2025): Intelligent lexicography. Proceedings of the eLex 2025 conference* (pp. 661–681). Lexical Computing CZ s.r.o.

Prys, D. & Jones, D. B. (2007). *Guidelines for the Standardization of Terminology for the Welsh Assembly Government Translation Service and the Welsh Language Board*. Welsh Language Board.

Vaus, M. & Janson, K. (2025). Consolidating terminology on a single online platform: The Estonian case. In R. Resi & F. Steurs (Eds.), *Handbook of Terminology: Volume 4. Terminology planning in Europe* (pp. 178–200). John Benjamins Publishing Company.

Zanola, M. T. & Resi, R. (2025). From terminological neology to terminology planning for corporate and professional initiatives: The Italian case. In R. Resi & F. Steurs (Eds.), *Handbook of Terminology: Volume 4. Terminology planning in Europe* (pp. 616–637). John Benjamins Publishing Company.

Contact information

Ana Ostroški Anić
Institute for the Croatian Language
aostrosk@ihjj.hr

Federica Vezzani
University of Padua
federica.vezzani@unipd.it

Kris Heylen
Instituut voor de Nederlandse Taal
kris.heylen@ivdnt.org

Ilan Kernerman
Lexicala by K Dictionaries
ilan@lexicala.com

Acknowledgements

This paper is based on work from COST Action ENEOLI, CA22126 (European Network on Lexical Innovation) supported by the European Cooperation in Science and Technology (COST), https://www.cost.eu/actions/CA22126/

Ana Ostroški Anić, Martina Pavić & Lidija Cvikić

TYPES OF DEFINITIONS IN CROATIAN PRIMARY SCHOOL SCIENCE TEXTBOOKS

Keywords Definitions; textbook definitions; specialized knowledge; specialized lexicography

From the first grades of primary education, children encounter abstract and complex concepts, many of which cannot be adequately represented by traditional or conventionally formulated definitions. Consequently, younger learners may struggle with understanding abstract concepts like energy, community, or climate. Previous studies have shown that children's ability to engage with scientific and abstract concepts is closely related to the form in which these concepts are presented, including sentence structure, vocabulary choice, and the type of context provided (Murray & Reuter, 2005; Gelman & Meyer, 2011). Research on concept acquisition further suggests that abstract concepts are acquired gradually and are initially grounded in emotionally relevant experiences and situations, particularly in early school years, age 6 to 9 (Caramelli, Setti & Maurizzi, 2004). Contemporary educational resources already reflect these findings, which is evident in different types of presentation of complex concepts in textbooks, particularly in the domain of science education, where the interdisciplinarity of concepts is more prominent than in other school subjects.

The aim of this study is to examine the types of definitions in Croatian primary school science textbooks, Grades 3 to 8, in order to analyze how key concepts are systematically presented to different age level learners. The data includes definitions and definitional explanations extracted from textbooks for the subjects *Nature and Society* (Grades 3 and 4), *Nature* (Grades 5 and 6), and *Biology* (Grades 7 and 8). The definitions analysed are part of a larger dataset of textbook definitions extracted for the purpose of creating an online, learner-oriented dictionary of scientific concepts for primary education. All textbooks have been officially approved and widely used in Croatian primary education. At the present stage, textbooks published by two major publishers, *Školska knjiga* and *Alfa*, are analysed. Textbooks published by a third approved publisher, *Profil Klett*, will be incorporated in the second stage of the study.

From each textbook, definitions or explanations of concepts highlighted in bold font were manually extracted. Currently, a total of 654 definitions and/or explanations have been independently annotated by two annotators trained in terminology management. The data has been annotated for a superordinate concept, if stated, salient concept characteristics, and conceptual relations. Inter-annotator agreement has not been measured yet, but a preliminary qualitative analysis based on 59 concepts, 45 definitions and 21 explanations (Ostroški Anić & Pavić, in press) showed an extremely high agreement of 97% in the annotation of superordinate concepts. A much lower agreement was found for identifying concept characteristics, 68%. The causes for lower agreement were identified to be avoided in further annotation process.

For this paper, definitions of the same concept were compared vertically across grade levels in order to identify differences in the type of definition (e.g., intensional, partitive, enumerative, descriptive, functional, etc.) and in concept characteristics (e.g., origin, composition, form, function, purpose, location, etc.). Qualitative analysis shows that concrete core science concepts like *water* and *organism* share rather a uniform approach to presenting the concepts in textbooks, resulting in widely accepted definitions, e.g. *Water is a liquid without colour, taste, or smell* – a variant of which was found in four out of six textbooks for Grades 3 and 4. All six

textbooks refer to the states of water, either within the definition itself or in accompanying explanations. The process of enriching the conceptual knowledge is evident from a Grade 5 definition: *Water is a good solvent for gases and mineral substances*, whereas a Grade 7 explanation of water further expands the scope of concept characteristics by placing water within the context of human physiological processes.

Variation in definitional structures observed during the analysis shows clear vertical enrichment of concept-related information presented in definitions, which mostly follows the cognitive development primary school children. Discrepancies, however, do exist in some textbooks, where graphic information is the primary source of understanding, and it is not always matched with appropriate linguistic context. The results of this study will be used for developing a prototype of an online, learner-oriented science dictionary for primary school students, which will offer multiple definitions aligned to educational level and conceptual depth. A resource of this kind would support learners' transition from experiential knowledge toward scientific conceptual network in line with similar approaches to knowledge description, e.g. the idea of flexible definitions (San Martín, 2022).

References

Caramelli, N., Setti, A., Maurizzi, D. (2004). Concrete and abstract concepts in school age children. *Psychology of Language and Communication*, 8(2), 19–34.

Gelman, S. A., Meyer, M. (2011). Child categorization. *Wiley Interdisciplinary Review: Cognitive Science*, 2(1), 95–105. https://doi.org/10.1002/wcs.96

Murray, C. G., Reuter, K. (2005). Children's acquisition of categories and the implications for research in the development of classification schemes. *ASIST*, 1–13.

Ostroški Anić, A., Pavić, M. (in press). Age-Dependent Conceptual Structure in Primary School Science Textbook Definitions. In F. Vezzani et al. (Eds.), *Proceedings of the 5th International Conference on Multilingual digital terminology today. Design, representation formats and management systems (MDTT 2026)*.

San Martín, A. (2022). A Flexible Approach to Terminological Definitions: Representing Thematic Variation. *International Journal of Lexicography*, 35(1), 53–74. https://doi.org/10.1093/ijl/ecab013

Contact information

Ana Ostroški Anić
Institute for the Croatian Language
aostrosk@ihjj.hr

Martina Pavić
Institute for the Croatian Language
mpavic@ihjj.hr

Lidija Cvikić
University of Zagreb, Faculty of Teacher Education
lidija.cvikic@ufzg.hr

Acknowledgements

This work was created as part of the project *Semantic Frames in the Croatian Language* (SEF) funded by the European Union – NextGenerationEU.

Andrej Perdih, Janoš Ježovnik & Dejan Gabrovšek

IMPROVING DICTIONARY HEADWORD LISTS WITH WORD-PREVALENCE METRICS

An Examination of Low-Prevalence, High-Frequency Words in Slovenian

Keywords Word prevalence; frequency; headword list; dictionary

Objective

This study evaluates whether word-prevalence data can improve the selection and review of dictionary headword candidates. It focuses on Slovenian lemmata that combine higher corpus frequency with lower prevalence, because this mismatch may signal problems that frequency alone does not reveal.

Background and data

Word prevalence is the percentage of a population that knows a word (Keuleers et al., 2015). Large online lexical-decision studies have produced prevalence norms for Dutch, English, Spanish, Catalan, and Italian, based on tens of thousands of words and large participant samples (Keuleers et al., 2015; Brysbaert et al., 2019; Aguasvivas et al., 2018, 2020; Guasch et al., 2023; Amenta et al., 2025). These studies show that very frequent words are generally widely known, whereas words at lower frequencies vary considerably in prevalence. Lew & Wolfer (2024) represents a rare example of applying word-prevalence data to the field of lexicography.

The ongoing Slovenian megastudy has so far collected one hundred YES/NO responses for each of 35,000 words from 39,063 participants in 48,659 sessions; the complete list contains 79,413 words. In each session, participants classify 120 letter strings as Slovenian words or nonwords. Among the 120 letter strings given in each session, 84 were real Slovenian words and 36 were not, The list was compiled from existing dictionaries (Perdih et al., 2025), which provide verified words (as opposed to potentially unverified corpus lemmata) and to serve as a motivational element at the end of a session – lists of successfully and unsuccessfully recognized words linked to the Fran Slovenian dictionary portal (https://www.fran.si) are provided to participants at the end of each session. Consequently, frequent corpus lemmata absent from the dictionaries are outside the present analysis.

Selection and classification

We selected the 50 highest-frequency lemmata in the Gigafida reference corpus of Slovenian (Krek et al., 2020) whose prevalence values were below zero, meaning that fewer than half of the participants recognized them. The least frequent item, *masen* 'pertaining to weight' occurred 832 times (0.62 per million tokens). We then classified the likely sources of the frequency–prevalence mismatch.

TABLE 1 Features of the 50 highest-frequency words known by fewer than half of the participants

Feature	No. of words
Lemmatization/tokenization error	15
Uncommon dictionary form	17
Terminology	8
Foreign	8
Interjection	2
Proper name homographs	2

Results

Lemmatization or **tokenization errors** account for 15 items (30%) and often inflate corpus frequencies. A further 17 items have **dictionary forms** that are **uncommon** in actual use: 13 adjectives, 3 verbs, and 1 pronoun. For example, the adjectival headword *kaven* 'pertaining to coffee' was recognized by 45.23% of respondents (prevalence value -0.118)[1] despite a frequency of 4,094 (3.07 per million tokens); its more frequent definite form *kavni* would probably be easier to recognize. Similarly, the infinitives *poiti* 'to run out' and *onemoči* 'to tire, to languish' are much less common than their inflected forms.

Other mismatches reflect restricted use or ambiguous status. Eight items are **terms**, including *obtežba* 'load, weight'; eight may be interpreted as **foreign**, although some are Slovenian loanwords, such as *go* 'Japanese game' and *song* 'a song genre', while others are now archaic or obsolete and only occur in the corpus as parts of foreign-language texts (*il* 'clay', but also Italian definite article). The two **interjections** *hi* and *jah* may be difficult to recognize as Slovenian words without context, and *hi* may also be associated with the English greeting. Two items are homographs of **proper names**. Some categories overlap: for example, *Li* may be lemmatized as *li*, while *li* may also result from incorrect tokenization of a masculine plural past-participle ending.

Implications for headword-list development

Word prevalence is useful as a diagnostic measure for high-frequency headword candidates. When combined with corpus frequency, it helps identify lemmatization and tokenization errors, specialized terms, foreign-language material, and items whose recognition depends strongly on context. These findings can support both the addition of new entries and the review or removal of existing entries: low prevalence may indicate corpus noise, opens a question of headword form user-friendliness, or suggests that a stylistic or terminological label may be required rather than simple exclusion of the headword.

References

Aguasvivas, J. A., Carreiras, M., Brysbaert, M., Mandera, P., Keuleers, E., & Duñabeitia, J. A. (2018). SPALEX: A Spanish Lexical Decision Database From a Massive Online Data Collection. *Frontiers in Psychology*, 9, 2156, 1–6. https://doi.org/10.3389/fpsyg.2018.02156

[1] The percentage values are typically converted using the inverse function of the standard normal distribution. As a result, prevalence values are usually expressed on a scale ranging approximately from –2.5 to +2.5, which is better suited to statistical analysis.

Aguasvivas, J., Carreiras, M., Brysbaert, M., Mandera, P., Keuleers, E., & Duñabeitia, J. A. (2020). How do Spanish speakers read words? Insights from a crowdsourced lexical decision megastudy. *Behavior Research Methods*, 52(5), 1867–1882. https://doi.org/10.3758/s13428-020-01357-9

Amenta, S., De Varda, A. G., Mandera, P., Keuleers, E., Brysbaert, M., & Marelli, M. (2025). The Italian Crowdsourcing Project: Visual word recognition times for 130,495 Italian words. *Behavior Research Methods*, 57(1), 26. https://doi.org/10.3758/s13428-024-02548-4

Brysbaert, M., Mandera, P., McCormick, S. F., & Keuleers, E. (2019). Word prevalence norms for 62,000 English lemmas. *Behavior Research Methods*, 51(2), 467–479. https://doi.org/10.3758/s13428-018-1077-9

Guasch, M., Boada, R., Duñabeitia, J. A., & Ferré, P. (2023). Prevalence norms for 40,777 Catalan words: An online megastudy of vocabulary size. *Behavior Research Methods*, 55(6), 3198–3217. https://doi.org/10.3758/s13428-022-01959-5

Keuleers, E., Stevens, M., Mandera, P., & Brysbaert, M. (2015). Word knowledge in the crowd: Measuring vocabulary size and word prevalence in a massive online experiment. *Quarterly Journal of Experimental Psychology*, 68(8), 1665–1692. https://doi.org/10.1080/17470218.2015.1022560

Krek, S., Arhar Holdt, Š., Erjavec, T., Čibej, J., Repar, A., Gantar, P., Ljubešić, N., Kosem, I., & Dobrovoljc, K. (2020). Gigafida 2.0: The reference corpus of written standard Slovene. In N. Calzolari (Ed.), *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)* (pp. 3340–3345). ELRA - European Language Resources Association. http://www.lrec-conf.org/proceedings/lrec2020/LREC-2020.pdf

Lew, R., & Wolfer, S. (2024). What Lexical Factors Drive Look-Ups in the English Wiktionary? *Sage Open,* 14(1). https://doi.org/10.1177/21582440231219101

Perdih, A., Gabrovšek, D., & Pavlič, M. (2025). Izdelava seznama besed za množično raziskavo razširjenosti slovenskih besed. *Slavistična revija*, 73(1), 121–138. https://doi.org/10.57589/srl.v73i1.4231

Contact information

Andrej Perdih
ZRC SAZU, Fran Ramovš Institute of the Slovenian Language
andrej.perdih@zrc-sazu.si

Janoš Ježovnik
ZRC SAZU, Fran Ramovš Institute of the Slovenian Language
janos.jezovnik@zrc-sazu.si

Dejan Gabrovšek
ZRC SAZU, Fran Ramovš Institute of the Slovenian Language
dejan.gabrovsek@zrc-sazu.si

Acknowledgements

This article was written as part of the research project Slovenian Word-Prevalence: An Online Mega-Study of Word Knowledge (J6-50199) funded by the Slovenian Research and Innovation Agency (ARIS).

Silvia Piccini, Andrea Bellandi, Giuliana Elizabeth Vilela Ruiz & Davide Saponaro

A SEMANTIC WEB MODEL FOR DIACHRONIC EXPLANATORY COMBINATORIAL DICTIONARIES

Capturing Conceptual Change and Denominational Variation

Keywords Explanatory Combinatorial Dictionary; diachronic lexicography; semantic web; OntoLex-Lemon; conceptual change; perdurantism; Hebrew marriage terminology

This paper presents a computational and Semantic Web–based model for Explanatory Combinatorial Dictionaries (ECDs), well-established lexicographic resources grounded in Explanatory Combinatorial Lexicology and Meaning–Text Theory (Mel'čuk et al., 1995). The model aims to make ECD structures machine-actionable, FAIR-compliant (Wilkinson et al., 2016), and interoperable with the Linguistic Linked Open Data ecosystem.

The proposal extends ECDs in two directions: by introducing diachronic depth into dictionaries usually conceived synchronically, and by modelling the phenomenon described in the literature as "denominational paths" (Aymerich et al., 2008), i.e. the ways in which lexical units foreground salient conceptual traits within and across languages.

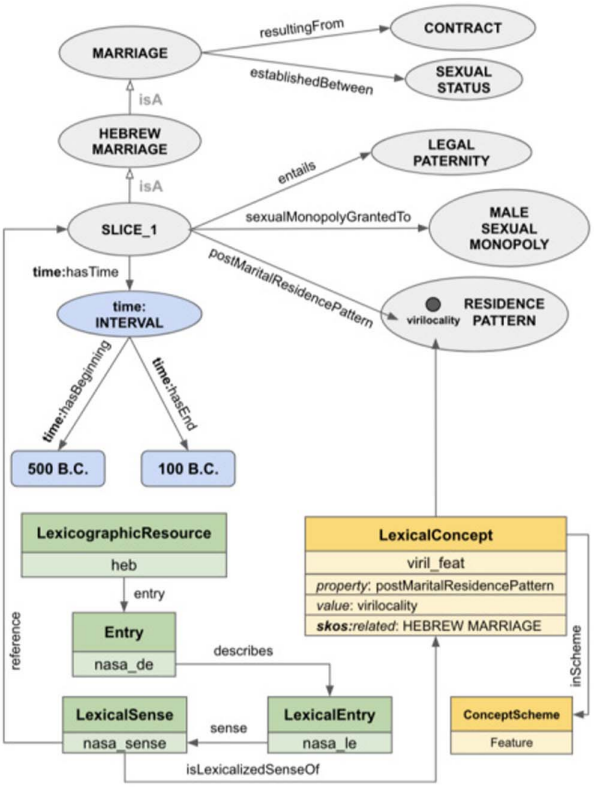
The proposal is illustrated through a computational diachronic ECD, currently under development, devoted to the terminology of marriage in ancient Israel. Marriage is a suitable test case because it is a polythetic concept whose conceptualisation varies across cultures and historical periods, including within the Jewish tradition (Piccini et al. 2024).

While Biblical sources (c. 500-100 BCE) represent marriage as a legal institution based on the acquisition of a woman and the transfer of jurisdiction from her *pater familias* to the husband—who acquires sexual monopoly and legal paternity—Talmudic sources (c. 150-500 CE) introduce a more sacralised conception, reflected in terminology foregrounding consecration. This conceptual polyhedrality is mirrored by several denominations, each lexicalising a salient trait: acquisition, as in לקח *laqaḥ* 'to take for oneself or for another'; possession, as in בעל *ba'al* 'to possess'; consecration, as in קידושין *qiddushin* 'consecrations'; and virilocality, as in נשא *nasa'* 'to lead', הושיב *hoshiv* 'to cause to dwell', and שלח *shillaḥ* 'to send'. The modelling of *nasa'* provides the illustrative example for the RDF/OWL-based representation presented in Figure 1.

The model combines a lexical and an ontological layer. The lexical layer relies on three vocabularies: OntoLex-Lemon[1] for lexical units and senses, Lexicog[2] for dictionary-like structure, and LexFom (Fonseca et al., 2016) for Lexical Functions. In Figure 1(b), *nasa'* is represented as an ECD entry specifying its canonical form, lexical sense and combinatorial information.

The ontological layer models diachronic conceptual change through a perdurantist, or 4D, approach (Welty et al., 2006), in which change is represented through temporal slices rather than by assigning time-dependent properties to the concept as a whole. Figure 1(a) presents the modelling of Hebrew marriage (`HEBREW_MARRIAGE`) between 500 and 100 BCE. This temporally bounded phase is represented by the class `SLICE_1`, defined as a subclass

[1] https://www.w3.org/2016/05/ontolex/
[2] https://www.w3.org/2019/09/lexicog/



(a)

```
01  :heb a lexicog:LexicographicResource .
02       dct:language "heb" .
03       lexicog:entry :nasa_de .

04  :nasa_de a lexicog:Entry .
05           :rdfs:label "nasa"@heb .
06           rdf:_1 :nasa_de_comp .
07           lexicog:describes :nasa_le .

08  :nasa_le a ontolex:Word ;
09           rdfs:label "nasa"@heb ;
09           lexinfo:partOfSpeech lexinfo:verb ;
11           ontolex:sense :nasa_sense .

12  :nasa_sense a ontolex:LexicalSense ;
13      lexfom:syn :laqaḥ_sense1 ;
14      lexfom:syn :baʿal_sense1 ;
15      lexfom:S1perf :ish_sense2 ;
16      lexfom:S2perf :ishà_sense2 ;
17      ontolex:reference :SLICE_1 ;
18      ontolex:isLexicalizedOf :viril_feat .

19  :nasa_de_comp a lexicog:LexicographicComponent ;
20      lexicog:describes :nasa_sense ;
21      rdfs:label "I" .

22  :SLICE_1 a owl:Class ;
23      rdfs:subClassOf :HEBREW_MARRIAGE ;
24      time:hasTime :slice_1_interval ;
25      :entails :LEGAL_PATERNITY ;
26      :sexualMonopolyGrantedTo
                  :MALE SEXUAL MONOPOLY ;
27      :postMaritalResidencePattern :virilocality .

28  :viril_feat a ontolex:LexicalConcept ;
29         skos:inScheme :Feature ;
30         skos:related :HEBREW_MARRIAGE ;
31         :property :postMaritalResidencePattern ;
32         :value :virilocality .
```

(b)

FIGURE 1 All figures (a) Ontological modelling of the concept of `HEBREW_MARRIAGE` according to a perdurantist (4D) approach, focusing on the Biblical time slice (c. 500–100 BCE) and its defining conceptual traits. (b) Turtle formalisation of the ECD entry נשא *nasa'*, illustrating the lexical layer (OntoLex-Lemon, Lexicog) and the encoding of Lexical Functions (LexFom)

of `HEBREW_MARRIAGE` and associated with a specific time interval. `SLICE_1` is characterised by traits such as legal paternity, male sexual monopoly and virilocal residence, each encoded as an existential restriction on a specific property. At the same time, it inherits more general, time-invariant traits from `MARRIAGE`, such as the establishment of a contractual bond and a socially regulated relation between sexes. Anchoring ECD entries to such slices the `ontoleox:reference` property turns the resource into a diachronic ECD.

The second objective, i.e. capturing the selective lexicalisation of salient conceptual traits, is implemented at the lexical–conceptual interface through the property `ontolex:isLexicalizedSenseOf`, which links a lexical sense (in our case *nasa'* lexical sense) to the conceptual trait it lexicalises (virilocality). Since this relation requires an `ontolex:LexicalConcept` as its object, conceptual traits are represented as lexical concepts within a feature-based schema, consisting of a `skos:ConceptScheme` and a set of features modelled as SKOS concepts (also typed as `ontolex:LexicalConcept`).

Such a formalisation supports fine-grained semantic and diachronic queries, from identifying the conceptual facet lexicalised by a term to tracing specific traits across historical phases (Figure 2).

Competency Questions	pseudo-SPARQL
Which properties characterise Hebrew marriage around 200 BCE?	**SELECT** ?property ?value **WHERE** { ?mar a :HEBREW_MARRIAGE ; time:hasTime [time:hasBeginning ?start ; time:hasEnd ?end] ; ?property ?value } **FILTER** (xsd:integer(?start) <= 200 && xsd:integer(?end) >= 200)
How does the concept of marriage evolve over time?	**SELECT** ?property ?value **WHERE** { ?mar a :HEBREW_MARRIAGE ; time:hasTime ?interval ; ?property ?value . ?interval time:hasBeginning ?start ; time:hasEnd ?end . } **GROUP BY** ?interval
Which lexical items denote Hebrew marriage around 200 BCE?	**SELECT** ?term **WHERE** { ?mar a :HEBREW_MARRIAGE ; time:hasTime [time:hasBeginning ?start ; time:hasEnd ?end] ; ontolex:isReferenceOf ?sense . ?le ontolex:sense ?sense ; rdfs:label ?term . } **FILTER** (xsd:integer(?start) <= 200 && xsd:integer(?end) >= 200)

Which terms lexicalise the trait of virilocality (and when)?	**SELECT DISTINCT** ?term ?start ?end
	WHERE {
	?sense ontolex:isLexicalizedSenseOf ?trait ;
	ontolex:reference [time:hasBeginning ?start ;
	time:hasEnd ?end] .
	?trait :property :postMaritalResidencePattern ;
	:value :virilocality .
	?entry ontolex:sense ?sense ;
	rdfs:label ?term .
	}

FIGURE 2 Example queries illustrating the expressive power of the proposed model

References

Aymerich, J. F., Fernández Silva, S. & Cabré Castellví, M. T. (2008). La multiplicité des chemins dénominatifs. *Meta*, 53(4), 731-747. https://doi.org/10.7202/019644ar

Fonseca, A., Sadat, F. & Lareau, F. (2016). Lexfom: a lexical functions ontology model. In M. Zock, A. Lenci & S. Evert (Eds.), *Proceedings of the 5th Workshop on Cognitive Aspects of the Lexicon (CogALex-V)* (pp. 145–155). The COLING 2016 Organizing Committee.

Mel'čuk, I. A., Clas, A. P. & Polguère, A. (1995). *Introduction à la lexicologie explicative et combinatoire*. Duculot- De Boeck.

Piccini, S., Saponaro, D. & Vilela Ruiz, G. E. (2024). La connotation logique en terminologie. Le mariage dans l'Israël antique comme étude de cas. *Cahiers de lexicologie*, 125(2), 103–129. https://doi.org/10.48611/isbn.978-2-406-17959-7.p.0103

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R. & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1), Article 160018.

Welty, C., Fikes, R. & Makarios, S. (2006). A reusable ontology for fluents in OWL. In B. Bennett & C. Fellbaum (Eds.), *Proceedings of the Fourth International Conference on Formal Ontology in Information Systems (FOIS 2006)* (Vol. 150, pp. 226–236). IOS Press.

Contact information

Silvia Piccini
Istituto di Linguistica Computazionale "A. Zampolli"
silvia.piccini@ilc.cnr.it

Andrea Bellandi
Istituto di Linguistica Computazionale "A. Zampolli"
andrea.bellandi@ilc.cnr.it

Davide Saponaro
Fondazione Rut
davide.saponaro@unint.eu

Giuliana Elizabeth Vilela Ruiz
Istituto di Linguistica Computazionale "A. Zampolli"
giulianaelizabeth.vilelaruiz@ilc.cnr.it

Acknowledgements

Scientific publication produced thanks to the agreement between the National Research Council – Institute of Computational Linguistics and the Rut Foundation.

Geraint Paul Rees, Blanca Arias-Badia, Elisenda Bernal, Marta Brescia-Zapata, Clara Anush Del Rey Adanalian, Ana Frankenberg-Garcia, Marta Garcia Casado, Paula Igareda González, Daniel Jiménez Casas, Robert Lew, Alba Milà Garcia, Guiomar Salvat, Aurora Troncoso Ruiz, Sergi Torner & Patrick Zabalbeascoa

ACCESSIBLE DIGITAL LANGUAGE RESOURCES (ADLAR)

Keywords Accessibility; inclusive lexicography; digital language reference resources; user evaluation; user studies; usability; digital exclusion; Easy-to-Understand (E2U) language; web accessibility; WCAG

A good deal of lexicographic research is concerned with evaluating or improving the accessibility of lexicographic data by making it easier for users to find and interpret. However, there is another sense of accessibility that has received less attention. This sense of accessibility concerns the design of resources so that they can be used by any person irrespective of personal features such as disability or impairment. Recent EURALEX congresses suggest a growing recognition of the importance of making lexicographic resources accessible in this sense. For instance, the XIX congress (in Alexandroupolis, Greece, 2021) was themed *Lexicography for Inclusion* while the XXI congress (in Cavtat, Croatia, 2024) featured a pre-conference workshop on *Lexicography and Accessibility* (Rees et al., 2024). Similarly, a handful of recent studies have demonstrated that existing dictionaries display a range of accessibility problems (Arias-Badia & Torner, 2023; Rees, 2023; Torres del Rey & García Garmendia, 2025).

However, in general, most research on the latter sense of accessibility in lexicography has suffered from several limitations. Firstly, the majority of studies have focused on a restricted set of users, mostly centring on persons with disabilities—especially visual impairments (Rees, 2025)—while the needs of those with cognitive or intellectual disabilities, as well as other disabilities, and more generally other vulnerable groups, have received little attention. Secondly, these studies have revealed little about how, or indeed if, users actually employ the lexical resources under evaluation. Thirdly, they have relied on compliance with guidelines such as the Web Content Accessibility Guidelines (WCAG) (World Wide Web Consortium (W3C), 2023)—a set of technical standards—as a proxy for accessibility evaluation rather than carrying out evaluation directly with users. Finally, existing research has dealt with monolingual resources for widely used languages, predominantly English and Spanish, while the accessibility needs of users of lower-resourced languages have not been directly addressed.

This poster outlines plans for the ADLaR (Accessible Digital Language Resources) project, which aims to address these gaps in previous research on accessibility in lexicography. Widening the range of participants considered, the project focuses on prototypical users who are at risk of digital exclusion, including persons with cognitive disabilities, many older adults—who remain among those "least likely to tap the potential of innovative technologies" worldwide (UN, n.d., para. 6)—and various sub-groups of migrants and refugees. These groups are among the target users of Easy-to-understand (E2U) texts, that is, texts produced using specific techniques intended to facilitate their comprehension by these populations (Jiménez-Andrés & Orero, 2022).

In the first, exploratory, phase of the project, a series of focus groups with users representing these cohorts is planned in order to establish whether their members make use of digital

language reference resources (DLRRs), which resources they use, and how they use them. It is envisaged that digital dictionaries, online language forums, machine translation, and AI chatbots may be among the resources mentioned. Following this, semi-automatic web accessibility evaluation (c.f. Rees, 2023) of the resources identified in the focus groups will also be undertaken in order to give an initial indication of their accessibility.

In the second, experimental, phase of the project, participants will be directly involved in evaluating the resources identified in the exploratory phase. In this second phase, a series of screen-recorded look-up and text production experiments is planned with the aim of assessing the extent to which DLRRs help users resolve language doubts. These will be followed by standardised usability testing conducted with the System Usability Scale (SUS) questionnaire (Brooke, 1996) to assess the usability (and by extension accessibility) of the DLRRs identified. Finally, follow-up interviews with a sample of participants will allow further insight into user experiences of each DLRR.

The results of these procedures will inform guidelines for the creation of accessible DLRRs. Most of the work will take place in Catalonia, allowing the investigation of DLRRs not only for English, a global lingua franca, and Spanish, a widely spoken global language, but also for Catalan, a language that has received scarce attention in research on accessibility in lexicographic resources.

Given the barriers to education, employment, and more generally to participation in society that exclusion from DLRRs could produce, it is important that E2U language users have access to suitable resources. It is hoped that the ADLaR project will go some way to fomenting such access. By presenting the project at EURALEX, the poster aims to invite feedback, foster collaboration, and engage the EURALEX community in the project.

References

Arias-Badia, B., & Torner, S. (2023). Bridging the gap between website accessibility and lexicography: Information access in online dictionaries. *Universal Access in the Information Society*, *23*, 545–560. https://doi.org/10.1007/s10209-023-01031-9

Brooke, J. (1996). SUS: A quick and dirty usability scale. In P. W. Jordan, B. Thomas, B. A. Weerdmeester, & I. L. McClelland (Eds.), *Usability evaluation in industry* (pp. 189–194). Taylor & Francis.

Jiménez-Andrés, M., & Orero, P. (2022). Digital multilingual practices in third-sector organizations. In S. O'Brian & F. M. Federici (Eds.), *Translating crises* (pp. 165–182). Bloomsbury Publishing Plc. https://doi.org/10.5040/9781350240117.ch-10

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. (2025). Making dictionary content accessible for people with visual impairments. *Lexikos*, *35*(2), 165–184. https://doi.org/10.5788/35-2-2077

Rees, G. P., Arias-Badia, B., Bernal, E., & Torner, S. (2024). Lexicography and accessibility. In K. Š. Despot, A. O. Anić, & I. Brač (Eds.), *Lexicography and semantics: Book of abstracts of the XXI EURALEX International Congress, 8–12 October 2024* (p. 290). Institute for the Croatian Language.

Torres del Rey, J., & García Garmendia, M. (2025). User interaction with assistive technology for a thorough evaluation of a WCAG 2-compliant e-dictionary: Assessing the accessibility of the Diccionario de la Lengua Española, v. 23.8. In I. Kosem, M. Jakubíček, M. Medveď, K. Zgaga, Š. Arhar Holdt, T. Munda, & A. Salgado (Eds.), *Electronic lexicography in the 21st century: Book of abstracts from the eLex 2025 conference. Intelligent lexicography* (pp. 31–33). Lexical Computing CZ s.r.o.

UN. (n.d.). *Why are digital skills critical for older persons?* https://social.desa.un.org/sdn/why-are-digital-skills-critical-for-older-persons

World Wide Web Consortium (W3C). (2023, October). *Web Content Accessibility Guidelines (WCAG) 2.2*. https://www.w3.org/TR/WCAG22/

Contact information

Geraint Paul Rees
Universitat Pompeu Fabra
geraintpaul.rees@upf.edu

Blanca Arias-Badia
Universitat Pompeu Fabra
blanca.arias@upf.edu

Elisenda Bernal
Universitat Pompeu Fabra
elisenda.bernal@upf.edu

Marta Brescia-Zapata
Universitat Pompeu Fabra
marta.brescia@upf.edu

Clara Anush Del Rey Adanalian
Universitat Pompeu Fabra
claraanush.delrey@upf.edu

Ana Frankenberg-Garcia
University of Surrey
a.frankenberg-garcia@surrey.ac.uk

Marta Garcia Casado
Universitat Pompeu Fabra
marta.garciac@upf.edu

Paula Igareda González
Universitat Pompeu Fabra
paula.igareda@upf.edu

Daniel Jiménez Casas
Universitat Pompeu Fabra
daniel.jimenezcasas@upf.edu

Robert Lew
Adam Mickiewicz University
rlew@amu.edu.pl

Alba Milà Garcia
Universitat Pompeu Fabra
alba.mila@upf.edu

Guiomar Salvat
Universitat Pompeu Fabra
guiomarailuo.salvat01@estudiant.upf.edu

Aurora Troncoso Ruiz
Radboud University Nijmegen
aurora.troncosoruiz@ru.nl

Sergi Torner
Universitat Pompeu Fabra
sergi.torner@upf.edu

Patrick Zabalbeascoa
Universitat Pompeu Fabra
patrick.zabalbeascoa@upf.edu

Acknowledgements

Grant ADLAR (ref. PID2024-157523OB-I00) funded by MICIU/AEI/10.13039/501100011033/ FEDER, UE

Randy Christian Saputra & Sílvia Barbosa

MAPPING EMERGING AI TERMINOLOGY THROUGH AUTOMATICALLY BUILT KNOWLEDGE GRAPHS

Keywords AI terminology; automatic knowledge graph construction; large language models; lexicography; neology; terminology extraction; glossary

The development of artificial intelligence (AI) and large language models (LLMs) has rapidly evolved new terminology such as *retrieval-augmented generation* (RAG), *context engineering*, and *model context protocol* (MCP), creating significant challenges for learners and researchers in tracking stable definitions and conceptual boundaries. MCP is a recent neologism in AI, introduced by Anthropic in 2024 to designate an open standard that enables LLMs to interact with external tools and data sources in a structured way. The term spread quickly as other AI providers adopted it, yet its definition and scope remain fluid across sources, positioning it in relation to concepts such as *function calling* and *agent frameworks*. MCP exemplifies how fast-moving technical innovation generates terminology whose meaning and relations are still stabilizing, highlighting the need for lexicographic resources capable of tracking emerging terms and their evolving semantic boundaries.

AI terms inside glossaries are often defined inconsistently, with unclear conceptual boundaries and loosely described relations. For instance, OpenAI's glossary defines "embeddings" as a vector representation preserving content meaning (OpenAI, n.d.), while Google's Machine Learning Glossary uses the term *embedding space* for the same concept (Google, n.d.). Therefore, a comprehensive AI glossary with clear lexical relationships is needed to help students and researchers track the latest developments.

Given the rapid expansion of AI neologisms, this study adopts explicit terminological principles grounded in ISO 704 (concept orientation, precision of conceptual relations, and definition by intension) and ISO 1087 (standardisation of terminological systems and conceptual modelling), operationalised in a knowledge graph (KG)-based lexicographic system. By integrating LLM-based extraction with these principles, the resource supports consistent treatment of new terminology, transparent definition writing, and traceable documentation across heterogeneous sources.

Despite rapidly evolving AI terminology, existing glossaries such as Google's Machine Learning Glossary are mostly static and lack explicit relational structures. This paper proposes a KG-based glossary of English-language AI terminology, integrating lexicographic principles with LLM-based extraction. KG-based terminology resources have proven effective in organizing specialized vocabulary from heterogeneous sources (Martín-Chozas, 2019; Nesterov et al., 2023), and contemporary lexicography is increasingly shifting from tree-structured entries to graph-based models that capture rich lexical-semantic relationships (Mechura, 2016). However, few studies have built domain-specific KGs for AI, and none have systematically utilized modern LLM-based extraction libraries such as LangExtract (Goel & Kiraly, 2025).

The proposed methodology integrates LLM-based extraction with a reproducible workflow: (i) corpus construction from arXiv papers, technical documentation, and curated AI glossaries; (ii) candidate term identification via domain filtering and pattern-based extraction; and (iii) definitional and relational extraction using LangExtract and structured prompts based on the SciDef framework (Kučera et al., 2026). The corpus is designed as a diachronic dataset,

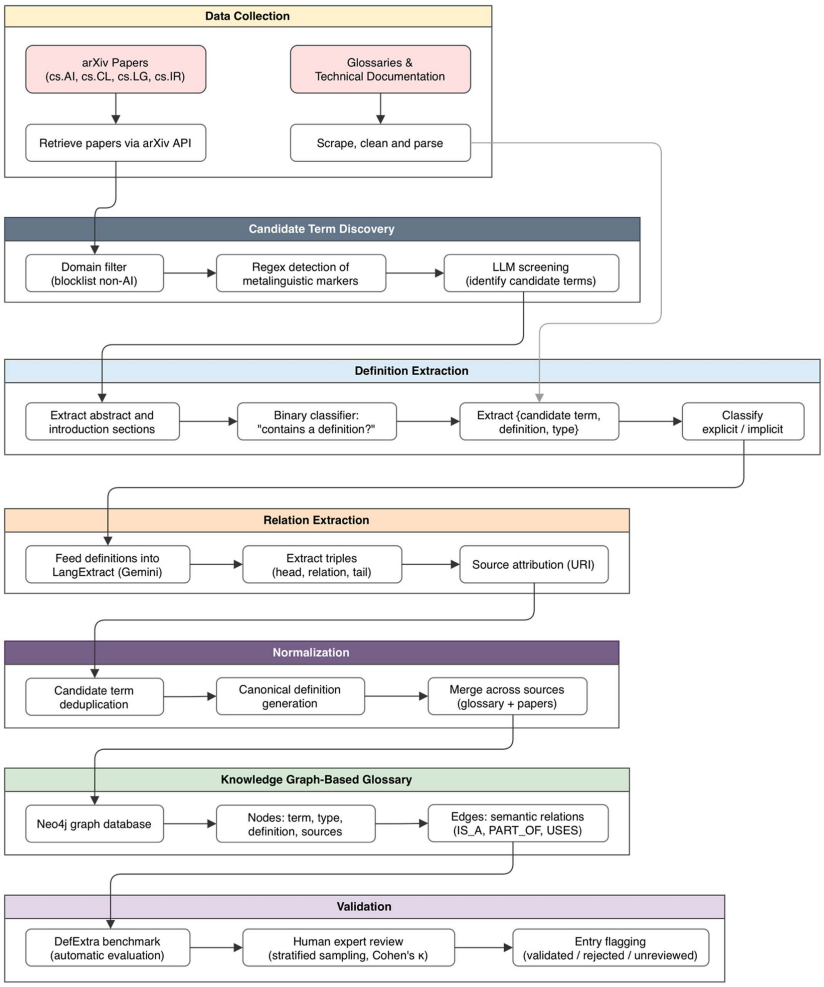


FIGURE 1 Overview of the KG construction pipeline that comprises seven stages: data collection, term discovery, definition extraction through multi-step LLM prompting, relation extraction, normalization, KG construction, and validation, using scientific papers and technical glossaries as primary data sources.

enabling the tracking of conceptual evolution through versioned term entries, allowing the KG to represent both synchronic and diachronic semantic variation. The extracted triples are stored in a graph database to form a KG-based glossary of AI terminology, operationalizing Alexiev's (2006) glossary model by organizing each term with its definition, synonyms, narrower terms, and exemplary contexts. Validation combines automatic evaluation against the DefExtra benchmark (Kučera et al., 2026) with independent assessment by two annotators with backgrounds in lexicography and computational linguistics. Inter-annotator agreement is measured using Cohen's kappa. Each entry is labelled as validated or unreviewed, enabling users to filter by confidence level.

The significance of this study has three aspects. First, this study proposes a methodology for constructing domain-specific lexicographic KG resources utilizing LLMs as extractors. Second, storing definitions and sources on graph nodes supports transparent terminology documentation and semantic search. Third, the reproducible workflow can be applied to other technical domains where terminological clarity and neologism detection are critical. Overall, this study presents a reproducible, lexicographically grounded framework for AI terminology, combining LLM-based extraction with terminological standards, structured validation, and KG-based modelling.

References

Alexiev, B. (2006). Terminology Structure for Learner's Glossaries. *Knowledge Organization*, *33*(2), 96–118. https://storage.imrpress.com/imr/journal/KO/article/505508/1752847386704.pdf

Goel, A. & Kiraly, A. (2025). *Introducing LangExtract: A Gemini powered information extraction library*. Google Developers Blog. Retrieved July 30, 2025, from https://developers.googleblog.com/en/introducing-langextract-a-gemini-powered-information-extraction-library

Google. (n.d.). *Machine Learning Glossary*. Retrieved November 25, 2025, from https://developers.google.com/machine-learning/glossary

Kučera, F., Mandl, C., Echizen, I., Timofte, R. & Spinde, T. (2026). *SciDef: Automating Definition Extraction from Academic Literature with Large Language Models.* (*arXiv preprint* No. *arXiv:2602.05413)*. arXiv. https://arxiv.org/abs/2602.05413

Martín-Chozas, P. (2019). Creation and enrichment of a Terminological Knowledge Graph in the legal domain. In *DC JURIX 2019: Seventh Doctoral Consortium of JURIX 2019* (pp. 1-11). CEUR Workshop Proceedings http://ceur-ws.org/Vol-2598/paper-07.pdf

Mechura, M. (2016). Data Structures in Lexicography: from Trees to Graphs. In A. Horák, P. Rychlý, & A. Rambousek (Eds.), *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2016* (pp. 97–104). Tribun EU. http://nlp.fi.muni.cz/raslan/2016/paper04-Mechura.pdf

Nesterov, A., Hollink, L., van Erp, M., & van Ossenbruggen, J. (2023). A Knowledge Graph of Contentious Terminology for Inclusive Representation of Cultural Heritage. In *The Semantic Web: 20th International Conference, ESWC 2023* (pp. 502–519). Springer-Verlag. https://doi.org/10.1007/978-3-031-33455-9_30

OpenAI. (n.d.). *Key concepts*. Retrieved November 27, 2025, from https://platform.openai.com/docs/concepts

Contact information

Randy Christian Saputra
NOVA FCSH
a2025135389@campus.fcsh.unl.pt

Sílvia Barbosa
NOVA FCSH CLUNL
silviabarbosa@fcsh.unl.pt

Acknowledgements

The research of Sílvia Barbosa is supported by the Portuguese national funding through the FCT – Portuguese Foundation for Science and Technology, I.P. as part of the project UID/03213/2025 – Linguistics Research Centre of NOVA University Lisbon (CLUNL) (https://doi.org/10.54499/UID/03213/2025).

Barbara Schmidt-Thieme, Theresa Kruse & Ulrich Heid

LEXICAL AND CONCEPTUAL TOOLS FOR LEARNING MATHEMATICS

Keywords Mathematics; learning; terminography; e-dictionaries; dictionary production

Studies with first-semester mathematics students indicate that after leaving highschool, they understand the subject primarily as a set of calculation rules, i.e. as a 'toolbox' (Törner & Grigutsch, 1994; Girnat & Hascher, 2021). However, mathematics as taught at university, as well as any scientific approach to mathematics, should rather be understood as an interconnected system of concepts expressed linguistically in a rich terminology. Learning mathematics therefore involves learning mathematical language and understanding how mathematical concepts are interrelated (cf. Riccomini et al., 2015).

Students are thus in need of lexical/terminological tools that support them in two ways: (a) to understand basic mathematical concepts and their relations (cognitive needs, in terms of the lexicographic function theory, cf. Tarp, 2008; Fuertes-Olivera & Tarp, 2014); and (b) to get acquainted with the phraseology needed for understanding and producing texts about mathematical topics (communicative needs: receptive and productive).

We analysed a number of sample entries from sources students may typically resort to: Wikipedia or other (online) encyclopedias, AI generated summaries of answers to keyword-based questions as well as (online) learning material for first years. We checked, a.o., the German terms 'Gleichung' (equation), 'Äquivalenz' (equivalence), 'Lösung' (solution) and 'Lösungsmenge' (solution set), all of which appear early in the first semester introductory lecture.

Wikipedia and similar tools provide substantial conceptual knowledge, but they require a considerable effort to extract relational concept data from the rather complex article texts (cf. Kruse, 2025, pp. 81-83). The analysed online material either contains wikipedia-like mini-essays or it focuses on computing examples. None of these make phraseology explicit.

We assume that formulating meaning explanations and conceptual relations explicitly may help students to master (and to memorize) mathematical concepts (a sort of 'learning by doing'). In an experimental course (cf. Kruse et al., 2024) administered since 2023, we thus invited participants of the first year introductory lectures to mathematics to restructure elements of the lecture contents by designing their own dictionary entries for the mathematical terms learned in the lecture; we also asked them to produce small ontological networks. To ensure that they would adhere to a general guideline, we gave a lexicographical mini-introduction to the microstructural items we are interested in, e.g. definitions, possibly paraphrases of definitions for school children, synonymy, hyp(er)onymy, related concepts that are part of the subdomain in question as well as collocations. We equally introduced the students to concept maps and ways to represent relations.

An analysis of a subset of the students' submissions (96 submissions) shows that their dictionary entries contain more types of microstructural items than their networks. All dictionary entries and almost 80 % of the networks contain definitions. Synonyms and Hyponyms are present in ca. 80 % of both types of submissions, while collocations only show up in dictionary entries and mathematical notations (symbols etc.) almost exclusively in the network graphs.

Overall, the students are quite successful in the exercise: they provide more conceptual relations than most online learning material. It seems, however, that they should need to be made

more aware of mathematical phraseology. A combination of particularized information in the style of dictionary entries and of small network-like representations may be a good way to provide them with their own learning material. In the medium term, a study of the effectiveness of the approach would be needed.

References

Fuertes-Olivera, P. A., & Tarp, S. (2014). *Theory and practice of specialised online dictionaries: Lexicography versus terminology*. De Gruyter.

Girnat, B., & Hascher, T. (2021). Beliefs von Schweizer Schülerinnen und Schülern zum konstruktivistischen und instruktivistischen Lernen im Mathematikunterricht der Sekundarstufe I – Ergebnisse eines Large-Scale-Assessments zur Überprüfung mathematischer Grundkompetenzen (ÜGK). *Unterrichtswissenschaft*, 525–546. https://doi.org/10.1007/s42010-021-00136-5

Kruse, T., Heid, U., Schmidt-Thieme, B. (2024). Mathematics Students as Lexicographers – Learning Domain Concepts and Their Relations by Designing Dictionary Articles and Concept Maps. In K. Š. Despot, A. Ostroški Anić & I. Brač, (Eds.), *Lexicography and semantics: Proceedings of the XXI EURALEX international congress* (pp. 711-719). Institute for the Croatian Language.

Riccomini, P.R., Smith, G.W., Hughes, J., Fries, K.M. (2015). The Language of Mathematics: The Importance of Teaching and Learning Mathematical Vocabulary. *Reading & Writing Quarterly*, 31(3), 235-252. https://doi.org/10.1080/10573569.2015.1030995

Törner, G., & Grigutsch, S. (1994). "Mathematische Weltbilder" bei Studienanfängern – eine Erhebung. *JMD,* 15, 211–251. https://doi.org/10.1007/BF03338808

Contact information

Ulrich Heid
Universität Hildesheim
heidul@uni-hildesheim.de

Theresa Kruse
Universität Hildesheim
kruse001@uni-hildesheim.de

Barbara Schmidt-Thieme
Universität Hildesheim
bschmidt-thieme@imai.uni-hildesheim.de

Nicolai Hartvig Sørensen & Andrea Stengaard

BEYOND WORD LISTS – BUT NOT BEYOND DICTIONARIES

Keywords Online dictionaries; user experience; UX design

The rapid rise of generative AI has prompted claims that traditional dictionaries, and perhaps even dictionary use itself, are becoming obsolete. We acknowledge that generative language models are transforming many aspects of information access and language-related tasks. However, we argue that they do not remove the need for high-quality, human-curated lexical resources. Generative systems, which rely on stochastic sampling and lack explicit provenance and systematic source control, cannot guarantee information that is fully accurate, transparent, and verifiable, and this seems unlikely to change. In this landscape, carefully edited lexical data remain not only relevant but essential. The challenge for lexicography is therefore not survival, but adaptation: competing for users' attention and integrating lexicographic strengths into an AI-saturated environment.

In this poster, we present the redesign of a large national dictionary portal that now brings together twelve dictionaries, all meaning-oriented rather than purely spelling-oriented. Our central claim is that online dictionaries have relied too long on opaque lists of word forms, especially in search results and autocompletion interfaces. Conventional interfaces typically display bare lemma forms, at most accompanied by a part-of-speech label. Such lists presuppose that users already know the words, or else force them to click through to each article before they can judge relevance. This is not only inefficient; it underuses one of lexicography's core assets: sense information.

We propose moving "beyond word lists" by integrating compact meaning information directly into all key navigational entry points. In the redesigned interface of our portal site, autocompletion results and other word lists are now always enriched with short snippets derived from dictionary definitions. Users thus see, from the outset, not only which headwords are attested in the dictionaries but also a concise indication of what each candidate means.

For instance, in our main contemporary dictionary, typing *malle* now automatically autocompletes with (cf. figure 1):

- *malle1 sb. "1. benfisk der ofte har skægtråde og en lille fedtfinne . . ."*
- *malle2 sb. "lille bøjle eller ring af metal der sammen med en tilsvarende hægte bruges til at lukke fx en vest e. . . ."*
- *mallemuk sb. "ca. 45 cm lang, mågelignende stormfugl med grå ryg og tykhals"*

This approach allows us to dispense with separate, formal disambiguation pages for homographs: instead of presenting a list of identical word forms that must be selected blindly, we display each homograph with an immediate sense-based cue while typing, ideally shortening the path from the query to the relevant entry.

A further design challenge concerns inflected forms. Many existing systems exclude inflected forms from autocompletion, often out of concern that the same lemma will appear repeatedly. In our case, the autocompletion list is intended to assume much of the work of a separate search results page, so excluding inflected forms would significantly weaken its usefulness. We address this by allowing exact matches on inflected forms while still providing the same

FIGURE 1 Typing *malle* in the search field

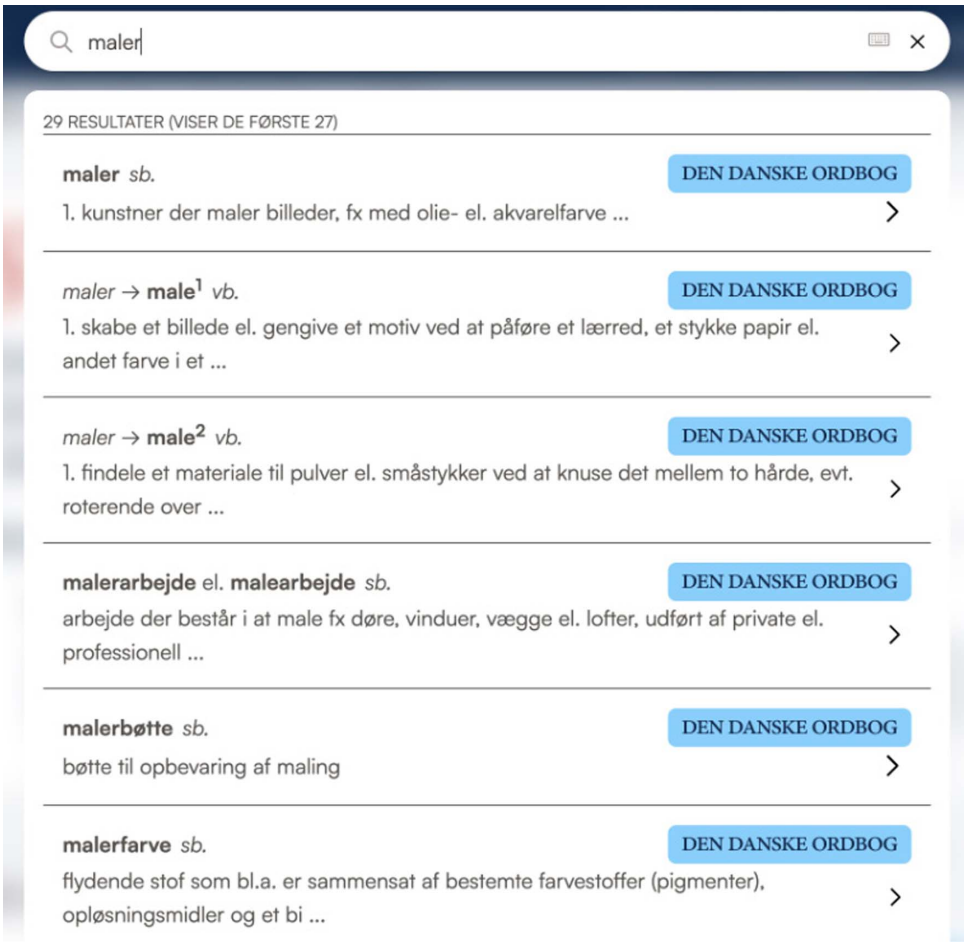


FIGURE 2 An autocompletion result while typing. Short snippets give the user hints about which lemma is the intended. Exact matches in inflected forms of *maler* is included (entry 2 and 3), but not inflected forms beginning with the typed string (e.g. existing inflected forms like *malere* and *malers*)

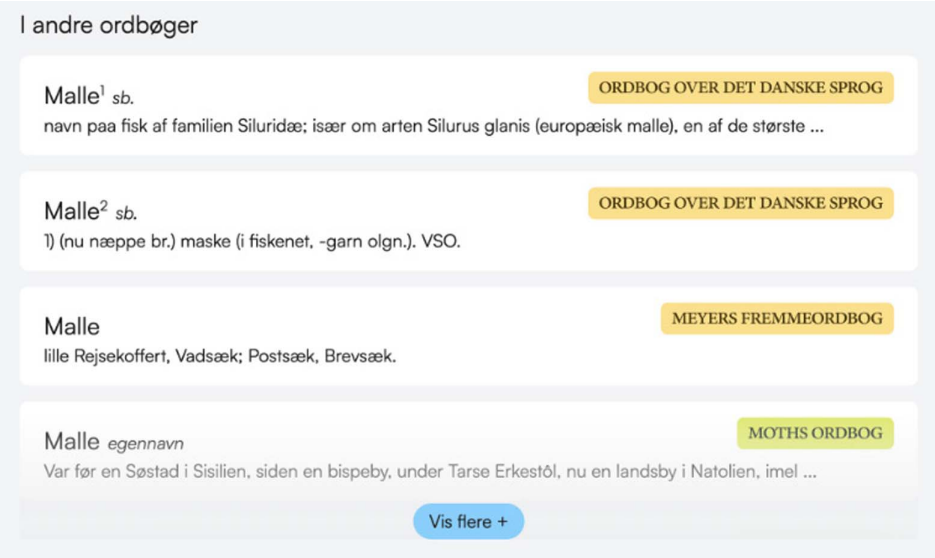


FIGURE 3 Links to other dictionaries than the current in the new platform. The user is able to determine which, if any, links might be relevant before clicking

definition snippets, enabling users to navigate efficiently from the forms they actually type to the relevant lexical entries, cf. figure 2.

In cases where the user types a string that does not correspond to a lemma in the dictionary, the autocompletion list starts showing words that similar in spelling to the string, i.e. a *Did you mean*? result.

Implementing this snippet strategy across a heterogeneous set of twelve dictionaries, including several retro-digitized works with only typographical markup and no explicit semantic structure, required developing methods for automatic snippet generation. In practice, we simply chose to present the first sense, or, in case of retro-digitized dictionaries, the beginning of the body of the entry, letting the sense number and the common abbreviation for ellipses, "...", suggest that the entry include more sense information than presented in the autocompletion. We argue that even these automatically derived, simplistic sense cues represent a substantial usability gain. Even these support more informed navigation, reduce fruitless clicks, and make the semantic richness of the dictionaries visible earlier in the user journey.

We also utilize these snippets when supplying links to matches of the current query in the 11 other dictionaries, cf. figure 3.

We contend that the future of dictionaries lies not in mimicking generative AI, but in foregrounding what lexicography uniquely provides: curated, structured, and interpretable information that guides users through the lexicon more intelligently than word lists alone ever can. In the future, when the new platform has been live for a sufficient time, we hope to have data to support this.

In this poster, we present the motivation for the new design, and in-depth examples of the queries, before and after the redesign, and illustrative search scenarios highlighting how snippets support cross-dictionary navigation.

Contact information

Nicolai Hartvig Sørensen
Det Danske Sprog- og Litteraturselskab
nhs@dsl.dk

Andrea Stengaard
Det Danske Sprog- og Litteraturselskab
as@dsl.dk

Stefania Spina, Irene Fioravanti & Fabio Zanda

THE LEARNER DICTIONARY OF ITALIAN COLLOCATIONS (DICI-A)

A New Lexicographic Resource

Keywords Collocations; learner dictionary; Italian

We introduce and describe the DICI-A, a learner dictionary of Italian collocations which is the output of a two-year project funded by the Italian Ministry of University and Research. In the Italian lexicographic landscape, three dictionaries of collocations have been published in the last fifteen years (Lo Cascio, 2013; Tiberii, 2012; Urzì, 2009). However, none of the three has two fundamental features that can be found in DICI-A: it is specifically targeted at L2 learners of Italian, and it has been created according to corpus-based criteria. Other features of the DICI-A are:

- it is monolingual;
- it includes over 10,000 collocation entries belonging to six syntactic configurations;
- each collocation is assigned to a specific proficiency label;
- GenAI and human assessment have been integrated for the creation of definitions and examples;
- it is digital and freely available online, at https://dictionary.dici-a.it/.

We relied on a broad definition of collocation: a co-occurrence of two words with a syntactic relation characterised by its conventional meaning, resulting from the number of times it is used in naturally occurring language (frequency), the range of texts where it occurs (dispersion) and the extent to which its components attract each other and are strongly associated (association measures), either adjacently or within a distance.

In addition to a general description, the presentation will focus on three specific features of the dictionary.

1) Identification and selection of the dictionary entries

We have included in the dictionary collocations that fall into six syntactic types: i. Verb + Direct object (vdobj; *mantenere una promessa*, 'to keep a promise'); ii. Adjective + Noun/ Noun + Adjective, the adjective is a modifier before or after a noun (amod; *brutta avventura*, 'bad adventure'; *tempo libero*, 'free time'); iii. Verb + Adjective, the adjective functions like an adverb by modifying the verb (advmod1; *stare zitto*, 'to stay quiet'); iv. Verb + Adverb, the adverb modifies the verb (advmod2; *fare presto*, 'to hurry up'); v. Adverb + Adjective, the adjective is modified by the adverb (advmod3; *altamente positivo*, 'highly positive'); and vi. Noun + Noun, compounds made of two adjacent nouns (comp; *parco divertimenti*, 'amusement park').The collocations belonging to these six syntactic configurations were extracted from the PEC24 (Spina et al., 2025), a large reference corpus of written and spoken Italian, by combining pos-tagging and syntactic parsing (Seretan, 2011). These initial 2 million candidate collocations were filtered via a multi-method approach (Spina et al., 2026) involving both automatic stages (based on five different measures: dispersion, frequency, mutual information, log-dice and log-likelihood) and human evaluation, after a comparison between the filtered candidate collocations and two existing non corpus-based Italian collocation dictionaries.

2) Attribution of proficiency labels to each entry

The 10,596 collocations resulting from this selection process were assigned CEFRCV-based proficiency labels (*Common European Framework of Reference Companion Volume*; Council of Europe, 2020) based on a set of quantitative and qualitative criteria, including corpus frequency and dispersion, semantic transparency, register, and CEFRCV vocabulary range descriptors. This procedure was entirely based on human assessment: each collocation was annotated independently by two annotators; then a third annotator resolved the cases of disagreement. Inter-annotator agreement varied across collocation types; as an example, it reached 80% for the verb-direct object collocations.

3) Creation of definitions and examples with the support of GenAI

We relied on recent literature (Lew, 2023; Lew et al., 2024) showing that GenAI can be effective in speeding up the process of dictionary creation, performing well both in writing definitions and in producing examples. Thus, we asked ChatGPT-4o to assist us in identifying collocation's meaning(s) and in providing us with definitions and examples for each meaning. In our prompt, we strongly emphasised the need for the output to contain learner-friendly vocabulary suitable for learners of Italian. The 10,596 AI-generated pairs of definitions and examples were validated by human lexicographers and replaced or modified accordingly. After this evaluation process 72% pairs could be accepted as they were in the AI-generated version; 12% required changes in the definition, 12% had to be modified in the example, and only 4% had to be completely rewritten.

References

Council of Europe (2020). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment – Companion Volume*. Council of Europe Publishing.

Lew, R. (2023). ChatGPT as a COBUILD lexicographer. *Humanities and Social Sciences Communications*, 10, 704. https://doi.org/10.1057/s41599-023-02119-6

Lew, R., Ptasznik, B., Wolfer, S. (2024). The effectiveness of ChatGPT as a lexical tool for English, compared with a bilingual dictionary and a monolingual learner's dictionary. *Humanities and Social Science Communication*, 11, 1324. https://doi.org/10.1057/s41599-024-03775-y

Lo Cascio, V. (2013). *Dizionario Combinatorio Italiano*. John Benjamins.

Seretan, V. (2011). *Syntax-based collocation extraction*. Springer.

Spina, S., Zanda F., Fioravanti I. (2025). From PEC to PEC24: A new reference corpus for Italian. *Italiano LinguaDue*, 17(1), 745–768. https://doi.org/10.54103/2037-3597/29101

Spina, S., Fioravanti, I., Zanda, F., Forti, L., Perri, D., Gervasi, O. (2026). Developing a learner dictionary of collocations: description and evaluation of a multi-method approach. *Corpus Linguistics and Linguistic Theory*, online first. https://doi.org/10.1515/cllt-2025-0008

Tiberii, P. (2012). *Dizionario delle collocazioni: Le combinazioni delle parole in italiano.* Zanichelli.

Urzì, F. (2009). *Dizionario delle combinazioni lessicali*. Convivium.

Contact information

Stefania Spina
Università per Stranieri di Perugia, Italy
stefania.spina@unistrapg.it

Irene Fioravanti
Università per Stranieri di Perugia, Italy
irene.fioravanti@unistrapg.it

Fabio Zanda
Sapienza Università di Roma, Italy
fabio.zanda@uniroma1.it

Petra Storjohann

DISCURSIVE DYNAMICS OF PEJORATION

The Term *Pushback* between Usage and Lexicography

Keywords Migration discourse; discursive shifts; emotive meaning; framing; pejoration

When a topic or lexical item is repeatedly invoked within public discourse in a negative context, such as through media framing, political rhetoric or social stereotypes, this can influence both the emotional responses and evaluative attitudes associated with the concept, thereby affecting the semantic prosody of the term as reflected in its connotations. Nevertheless, its core denotative meaning typically remains stable over time. This well-established discursive mechanism has been extensively discussed in introductory linguistics texts and semantics handbooks (Blank, 1999; Traugott, 2006; Geeraerts, 2010). However, the underlying process of connotational development unfolds gradually and convincing empirical demonstrations of processual shifts rather than the pure results within the discursive environment are scarce (e.g. Baker, Vessey & McEnery, 2021). Moreover, the trajectory of such lexical transformations including the development of emotive meaning is rarely or inadequately documented in lexicographic sources (cf. Atkins & Rundell, 2008; Stubbs, 2001).

This paper pursues three objectives. Firstly, it examines the evolving sentiment associated with the German term *Pushback*, an English borrowing that entered the language around 2015 within the broader migration discourse. Over the subsequent decade, the discursive context surrounding the term has evolved, leading to a marked increase in the affective intensity of its contemporary usage. This pejorative shift, particularly in political discourse, is driven by repeated stylistic devices, invoking associations, framing effects, attitude indexing and the gradual establishment of semantic prosody, all of which can be traced diachronically (Chilton, 2004). Particular attention is paid to shifting features such as semantic prosody, affective meaning, stance-taking, politicisation, and legitimisation, all of which have become progressively entrenched in the neologism over the past decade. To trace the development of these features over time, the study employs a diachronic corpus-based analysis structured around sequential time slices. Secondly, the paper proposes a method for incorporating these findings within a concrete lexicographic entry, aiming to sensitise users to discursive shifts, semantic developments and language change more generally. It further seeks to provide dictionary information that capture the evolving contextual environment, making communicative strategies, framing, associative patterns and evaluative contexts observable over time. This necessitates the adoption of innovative entry structures and lexicographic presentation modes that can be integrated into a new neologism resource (Storjohann, 2024). However, the structure and design must remain flexible and capable of adapting to the diachronic and discursive characteristics of the headwords. The new resource permits an open range of descriptive approaches and representational formats (see Storjohann forthcoming). Such flexible options allow for a lexicographic treatment of controversial neologisms engaging systematically with numerous discursive dimensions in order to capture adequately the meaning and the development of its corresponding discourse. The integration of *Pushback* as a diachronic descriptive dictionary entry serves as an illustrative example of potential future developments where specific new lexical and discursive phenomena are given extra space and focus. Finally, I will demonstrate how AI can be effectively employed to support lexicographic practice at multiple stages. These include compiling supplementary data, collecting relevant historical information to assess the impact of political decisions on language use and facilitating the visualisation of linguistic analyses.

References

Atkins, S. B. T. & Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford University Press.

Baker, P., Vessey, R. & McEnery, T. (2021). *The Language of Violent Jihad*. Cambridge University Press.

Blank, A. (1999). *Why do new meanings occur? A cognitive typology of the motivations for lexical semantic change*. In A. Blank & P. Koch (Eds.), *Historical Semantics and Cognition* (pp. 61–90). De Gruyter.

Chilton, P. (2004). *Analysing Political Discourse: Theory and Practice*. Routledge.

Geeraerts, D. (2010). *Theories of Lexical Semantics*. Oxford University Press.

Storjohann, P. (2024). IDS-*Neo*$^{2020+}$: A Novel Resource for New German Words in Use. *International Journal of Lexicography*, 37(4), 389-403.

Storjohann, P. (forthcoming). The interplay between individual new lexical items and multi-layered discourse representations of neologisms. In Ph. Stöckle, W. Koppensteiner, K. Korecky-Kröll, M. L. Krapp, & T. Wissik (Eds.), *Lexicography in the Age of AI. Proceedings of the XXII EURALEX International Congress*. Austrian Academy of Sciences Press.

Stubbs, M. (2001). *Words and Phrases: Corpus studies of lexical semantics*. Blackwell Publishers.

Traugott, E. C. (2006). *Semantic change: Bleaching, strengthening, narrowing, extension*. In K. S. Goodman & Y. M. Goodman (Eds.), *Encyclopedia of Language & Linguistics* (pp. 124–131). Elsevier.

Contact information

Petra Storjohann
Leibniz-Institut für Deutsche Sprache, Mannheim
storjohann@ids-mannheim.de

Maria Tuulik, Kristina Koppel, Ene Vainik, Margit Langemets, Lydia Risberg, Hanna Maask, Esta Prangel & Liina Lutsepp

FROM USER FEEDBACK TO BETTER DEFINITIONS

Applying LLMs for Pedagogical Lexicography

Keywords Pedagogical lexicography; LLM-s; Estonian language; user research

In the presentation we will introduce the research on applying LLMs to help create definitions for learners of Estonian as a second language. For Estonian, there have been experiments conducted to create general language dictionary definitions with LLMs (see Tuulik et al., 2025). For English, LLM-s are tested for generating learner's definitions as well (see for example Ide et al., 2026; Rees & Lew, 2024). Our research is based on a survey we conducted with 55 learners representing different proficiency levels (A1–C1). The aim of the survey was to investigate how learners of Estonian comprehend dictionary definitions in a learner-oriented digital resource and identify linguistic features that support or hinder pedagogical effectiveness. Participants evaluated two alternative definitions – an original and a pedagogically adapted version – for 20 headwords from a dictionary portal *Learner's Sõnaveeb*. The analysis focused on lexical, morphological, and syntactic parameters influencing comprehension.

The results showed that learners consistently prefer pedagogically simplified definitions that are compact, clearly structured, and free of semantically empty pronouns, excessive conjunctions, and abstract modifiers. At lower proficiency levels (A2–B1), learners tend to favour familiar and internationally transparent loanwords, while this preference diminishes at higher levels. Morphological analysis revealed that the impersonal voice does not significantly impede comprehension, whereas non-finite constructions present notable difficulties. From a syntactic perspective, comprehension is enhanced when the headword appears at the beginning of the definition and when sufficient, relevant information is provided, even at the expense of brevity.

We are applying the results of the survey to adjust the existing definitions in *Learner's Sõnaveeb* and to guide LLMs to create definitions for new B2-level headwords. Based on expert evaluations, the best performing model is selected for the task. We prompt LLMs using new principles and exemplary definitions to explore how to automatically get best quality draft definitions for lexicographers to edit. Using drafts provided by LLMs would significantly speed up the compilation process and thus allow for a faster expansion of the headword set in *Learner's Sõnaveeb*.

References

Ide, Y., Nohejl, A., Tanner, J., Yanaka, H., Lindsay, C. & Watanabe, T. (2026). *Towards Automated Lexicography: Generating and Evaluating Definitions for Learner's Dictionaries*. arXiv. https://arxiv.org/abs/2601.01842

Keeleõppija Sõnaveeb. Eesti Keele Instituut. https://sonaveeb.ee/lite

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, 50-74. https://doi.org/10.1093/ijl/ecad030

Tuulik, M., Vainik, E., Prangel, E., Langemets, M., Aedmaa, E., Koppel, K. & Risberg, L. (2025). Tähenduste seletamine leksikograafias: kuivõrd on abi suurtest keelemudelitest? [Describing senses for

lexicography: how helpful are large language models?]. *Eesti ja soome-ugri keeleteaduse ajakiri [Journal of Estonian and Finno-Ugric Linguistics]*, 16(2), 147-176. https://doi.org/10.12697/jeful.2025.16.2.05

Contact information

Maria Tuulik
Institute of the Estonian Language
maria.tuulik@eki.ee

Kristina Koppel
Institute of the Estonian Language
kristina.koppel@eki.ee

Ene Vainik
Institute of the Estonian Language
ene.Vainik@eki.ee

Margit Langemets
Institute of the Estonian Language
margit.langemets@eki.ee

Lydia Risberg
Institute of the Estonian Language
lydia.risberg@gmail.com

Hanna Maask
Institute of the Estonian Language
hanna.maask@eki.ee

Esta Prangel
Institute of the Estonian Language
esta.prangel@eki.ee

Liina Lutsepp
University of Tallinn
llutsepp@tlu.ee

Efi Tzelepi & Zoe Gavriilidou

DIGITAL LEXICOGRAPHY FOR STEM EDUCATION

Designing a Multilingual Digital Glossary of Scientific Terminology in Greek Sign Language

Keywords Digital lexicography; Greek Sign Language; STEM terminology; multilingual glossary; inclusive education

Despite increasing efforts toward inclusive education, Deaf and hard-of-hearing students continue to encounter barriers in STEM education, mainly due to the limited availability of systematically documented scientific terminology in Greek Sign Language (GSL). This lack of accessible terminological resources restricts conceptual understanding, classroom participation and engagement with scientific knowledge. At the same time, specialised STEM educational resources require multilingual terminological alignment, concept-oriented organisation and accessible multimodal representation (Cabré, 1999; Pavel & Nolet, 2001).

In this study, STEM is used as an umbrella term covering Science, Technology, Engineering and Mathematics, while Natural Sciences are treated as a subcategory within the Sciences. The glossary focuses primarily on terminology from the Natural Sciences included in secondary education STEM curricula.

The paper presents the methodological design and development of a multilingual digital glossary of STEM terminology in GSL. The glossary integrates Greek, English and German scientific terminology with multimodal GSL representations and accessibility-oriented digital lexicographic design, drawing on principles from digital lexicography, terminology studies and inclusive educational design (Jackson, 2013; Gavriilidou & Mavromatidou, 2016).

Greek functions as the language of schooling, English as the dominant language of international scientific communication and German as an additional scientific language with historical importance in European scientific terminology, especially in the Natural and Earth Sciences (Howarth, 2020; Sfoini, 2021).

The glossary adopts a multilingual and multimodal structure in which verbal terminological equivalents are connected with visual and linguistic representations in GSL. Scientific concepts are represented through written terminology, visual support and GSL video representations, enabling accessibility across different linguistic and sensory modalities. These representations include recorded GSL signs, visual conceptual support and pedagogically adapted explanatory material designed to facilitate conceptual accessibility for Deaf learners (Charamba, 2019, 2020).

Methodologically, the glossary development followed a five-phase protocol grounded in contemporary approaches to digital lexicography and terminology management (Klosa, 2013). First, a specialised school corpus was compiled from Greek secondary education Geology-Geography textbooks and analysed using AntConc, processing 18,206 lexical units. Through term extraction and evaluation procedures, 448 core scientific terms and 155 supplementary terms were identified.

The terminology set subsequently underwent empirical multi-criteria evaluation according to six parameters: curricular relevance, scientific validity, corpus frequency, conceptual

transparency for learners, representability in GSL and pedagogical accessibility. Quantitative evaluation was supplemented by qualitative feedback from experts in STEM education, GSL and lexicography, allowing iterative refinement of the glossary entries.

The fourth phase involved the lexicographic and terminological structuring of the resource. The macrostructure combines thematic and alphabetical organisation to support efficient retrieval and conceptual navigation. At the microstructural level, each entry integrates multilingual terminological equivalents, educational definitions, conceptual cross-references and multimodal GSL representations informed by the linguistic properties of GSL as a natural language (Sapountzaki, 2015; Sapountzaki et al., 2025).

The final phase focused on accessibility-oriented digital implementation. The platform incorporates high-contrast visual design, adaptable typography, linear navigation and compatibility with assistive technologies, aiming to ensure usability for learners with diverse sensory profiles, including deafblind users. Future validation procedures involving Deaf learners, educators and GSL specialists are also planned in order to evaluate the comprehensibility and pedagogical effectiveness of the glossary.

Overall, the paper contributes to current discussions on inclusive STEM education, multilingual terminology management and digital lexicography by proposing a systematically designed educational glossary integrating multilingual scientific terminology with multimodal sign language representation. The study offers a methodological framework that may inform the development of similar accessible terminological resources in other sign language and educational contexts.

References

Anthony, L. (2024). *AntConc* (Version 4.3.1) [Computer software]. Waseda University. https://www.laurenceanthony.net/software/AntConc

Cabré, M. T. (1999). *Terminology: Theory, methods, and applications* (Vol. 1). John Benjamins Publishing.

Charamba, E. (2020). Translanguaging in a multilingual class: A study of the relation between students' languages and epistemological access in science. *International Journal of Science Education*, 42(11), 1779–1798. https://doi.org/10.1080/09500693.2020.1783019

Charamba, E. (2021). Learning and language: towards a reconceptualization of their mutual interdependences in a multilingual science class. *Journal of Multilingual and Multicultural Development*, 42(6), 503–521. https://doi.org/10.1080/01434632.2019.1707837

Gavriilidou, Z., & Mavromatidou, S. (2016). Construction of tool for the identification of electronic dictionary users' skills. In T. Margalitadze & G. Meladze. (Eds.), *Proceedings of the XVII EURALEX International Congress. Lexicography and Linguistic Diversity* (pp. 168–178).

Howarth, R. J. (2020). Etymology in the Earth Sciences: From "geologia" to "geoscience". *Earth Sciences History*, 39(1), 1–27.

Jackson, H. (2013). *Lexicography: An introduction* (2nd ed.). Routledge.

Klosa, A. (2013). The lexicographical process. In R. Gouws, U. Heid, W. Schweickard & H. E. Wiegand (Eds.), *Dictionaries: An international encyclopedia of lexicography* (pp. 517–524). Walter de Gruyter.

Pavel, S., & Nolet, D. (2001). Handbook of terminology. Public Works and Government Services Canada.

Sapountzaki, G. (2015). Greek Sign Language. In J. Bakken Jepsen, G. De Clerck, S. Lutalo-Kiingi & W. B. McGregor (Eds.), *Sign languages of the world: A comparative handbook* (pp. 311–336). De Gruyter Mouton.

Sapountzaki, G., Chaleftira, K., & Rontogianni, A. (2025). *Brief description of Greek Sign Language*. Kallipos – Open Academic Editions.

Sfoini, A. (2021). *Scientific terminology and translation during the Neo-Hellenic Enlightenment*. Almagest.

Contact information

Efi Tzelepi
Democritus University of Thrace
etzelepi@helit.duth.gr

Zoe Gavriilidou
Democritus University of Thrace
zoegabriil@gmail.com

Perina Vukša Nahod

FROM FIELD RECORDINGS TO AN ONLINE DICTIONARY

A Model for Processing Štokavian Dialect Material

Keywords Croatian dialect lexicography; Štokavian dialect; Neo-Štokavian Ijekavian; Neo-Štokavian Ikavian; online dialectological dictionary

This poster introduces a multi-modular model for developing an online Croatian dialect dictionary. Although Croatian dialect lexicography advanced significantly in the 20th century (Lisac, 2006) and continues to develop in the 21st century, online Croatian dialect dictionaries remain uncommon. Most available resources are digital reproductions of print dictionaries or amateur compilations that lack adherence to established scientific principles of dialectological analysis.

The Croatian language exhibits a threefold dialectal division. However, no existing dictionary encompasses material from all three Croatian dialects, nor does any cover more than one dialect (Magaš, 2026). Additionally, there is no Croatian dialect dictionary with a multi-modular structure similar to that developed for the Croatian standard language (Hudeček & Mihaljević, 2024). Despite the predominance of Štokavian speakers in Croatia, dictionaries dedicated to individual Štokavian local varieties are limited.

To address this gap, the project focuses on Štokavian dialects and aims to develop a dialect dictionary that integrates multiple corpora, various levels of linguistic processing, and diverse target age groups. This dictionary is intended to serve as a model for processing the lexical material of other Croatian local varieties, groups of varieties, and dialects, especially those recognized as intangible cultural heritage.

Within the project *Rijekom neretvanskih riječi – od kuće do škole* ("Along the River of Neretva Words – From Home to School"), an online dialectological Dictionary of Neretva Dialects is under development. This dictionary includes varieties from two Štokavian dialects: Neo-Štokavian Ijekavian and Neo-Štokavian Ikavian. It represents the first attempt to present material from two Croatian dialects simultaneously and to implement transparent, functional lexicographic solutions. The dictionary features modules designed for pupils from the first grade of primary school through the fourth grade of secondary school. It is also the first Croatian online dialect dictionary to incorporate multiple fields, including headword, linguistic annotation (expandable), meaning, sentence example, idioms, audio recording, video recording, drawing, additional notes, and references.

Because dialect dictionaries frequently omit forms essential for precise phonological and morphological description (Kapović, 2008), this project emphasizes the recording of grammatical forms relevant to determining the accentual paradigms of individual lexemes. Dictionary entries are compiled using TshwaneLex software, with three distinct modules developed:

1. a module for grades 1–4 of primary school, which includes accented lemmas, meanings, drawings, additional notes, audio recordings, and video recordings;
2. a module for grades 5–8 of primary school, which includes accented lemmas, grammatical forms, meanings, and audio recordings;

3. a secondary-school module, which includes accented lemmas, grammatical forms, extended meanings, idioms, accented sentence examples, and audio recordings.

The core corpus of the dictionary comprises lexical items collected through a dialect survey questionnaire that covers a range of thematic domains, including family relations, fruits and vegetables, agricultural work, life by the river, children's games, recipes, and customs. To prevent restriction to predefined thematic fields, the corpus is systematically expanded with lexical items extracted from recordings of informants' spontaneous speech. This thematic approach to data collection is particularly appropriate for primary school pupils in grades 1–4 and serves as a basis for further corpus expansion. Pupils in grades 5–8 actively participate in collecting and processing additional material, thereby contributing to both the educational and research aspects of the project.

Because the corpus includes material from two dialects, the dictionary records different phonological and morphological variants of lexical items. Each variant is accompanied by an abbreviation of the settlement where it was documented, enabling precise representation of local variation and systematic lexicographic treatment.

Secondary school students collect phraseological material using a targeted questionnaire developed from previous research on Neo-Štokavian dialects and through the investigation of selected concepts, such as strength, beauty, and stupidity. These concepts reveal conventional perceptions, stereotypes, and value systems characteristic of the speakers of a particular community. This approach enables systematic documentation of phraseological units and provides insights into the cultural and cognitive dimensions of dialectal language use.

By presenting these dictionary structures, the project offers insight into a contemporary and standardized approach to processing Croatian dialect material, ensuring both accessibility and searchability.

References

Hudeček, L. & Mihaljević, M. (2024). Hrvatski mrežni rječnik – Mrežnik. In L. Hudeček, M. Mihaljević & Ž. Jozić (Eds.), *Anatomija rječnika: Hrvatski mrežni rječnik* (pp. 2–34). Institut za hrvatski jezik.

Kapović, M. (2008). Ustrojstvo rječničkoga članka u dijalektnim rječnicima s obzirom na naglasak. *Hrvatski dijalektološki zbornik*, 14, 197–205.

Lisac, J. (2006). Hrvatska dijalektalna leksikografija u 20. stoljeću. In J. Hekman (Ed.), *Hrvatski jezik u XX. stoljeću* (pp. 225–237). Matica hrvatska.

Magaš, I. (2026). About the authors, motivations, and perspective in Croatian dialectal lexicography. *Croatica et Slavica Iadertina*, 21(2), 415–429.

Contact information

Perina Vukša Nahod
Institute of Croatian Language
pvuksa@ihjj.hr

Jonas Winnerlöv

LEARNER'S DICTIONARIES BASED ON ETYMOLOGICAL PRINCIPLES

The Example of a German Textbook Dictionary for Swedish Learners

Keywords Learner's dictionary; vocabulary learning; etymology; word family; root family; cognate; intralingual/interlingual facilitating factors

Today there are numerous frequency dictionaries marketed as learner's dictionaries, with various source languages and mostly with English as target language. They typically contain 5000 lemmas, which happens to coincide with a rule of thumb (though without much academic substance) that this number of words provides the lexical text coverage (95-98%) necessary for an adequate text comprehension.

For a learner of a foreign language it must be a daunting task to learn so many words in frequency or alphabetical order, or rearranged in thematic groups, as is often the case for learner's dictionaries. From a receptive perspective, it is however not necessary to learn all of them as vocabulary items, thanks to intralingual and interlingual facilitating factors.

Intralingual aspects: In a compounding language like German, complex words are more or less transparent and can be interpreted from their component affixes and root morphemes. The concept of "word family" in frequency studies usually means words in different parts of speech transparently derived from each other, for example *information, inform, informative*. Word family can also denote groups of words with a deeper morphological relationship, such as presented in the German "Wortfamilienwörterbuch" by August (2009), for instance *Mund: Mundwinkel, Mundart, mundfaul, mündlich, münden, Mündung, ausmünden, einmünden*, etc. To distinguish these deeper word formation families from the more superficial word family, the term used here is "root family" when etymologically related words are grouped around a primary word. A "primary word" is a word with one lexical root and at the same time the most central in its word and root family, such as *Mund/mouth*.

Interlingual aspects: Different language pairs, such as French–English or English–Finnish, have a varying proportion of cognates, understood here as any two words with an etymological link. A structural analysis of all the primary words in a German frequency dictionary (Tschirner & Möhring 2020) shows that over 90% of them has a distinguishable Swedish cognate. A study conducted with adult beginning learners of German at Stockholm University (Winnerlöv 2013) suggests that 70% of the primary content words can be understood thanks to their Swedish (or sometimes English) cognate. The rest of the cognates follow certain patterns and could be deciphered with some instruction, and when the patterns are less clear, they could at least serve as mnemonic help.

The new etymology-based German–Swedish learner's dictionary proposed here incorporates the above aspects and covers the majority of the words from the analysed frequency dictionary as well as some other words. During the presentation samples will be shown (where relevant with English rather than Swedish translations) from its main chapters, consisting of 1) some full cognates for pronunciation practices 2) cognate pairs according to predictable patterns 3) other more diverging cognates 4) non-cognate primary words with 1-1 translation equivalents 5) analytical articles for false friends, polysemic words, and other words that need

comments 6) function words in paradigms 7) other words learnt in semantic paradigms or series, and finally 8) root family groups. Since the dictionary is meant to be read in its entirety, and reviewed repeatedly, it is called a "textbook dictionary".

The dictionary is conceived to serve as a model for learner's monodirectional textbook dictionaries for other language pairs. The relative size of the intralingual versus the interlingual parts will depend on the structural distance between the languages in question, and also on the extent to which other commonly known languages such as English come into play. Common for all possible language versions is the design to reduce the relevant word population to a minimum number of primary words that need to be memorized as traditional vocabulary items, while relying on other cognitive strategies for grasping the meaning of the remaining words.

While the psycholinguistic aspects of cognates are well researched, there are still few scientific studies on their practical use in language pedagogy, which may reflect the traditional reluctance of language instructors to make the etymological dimension explicit, especially for younger and less advanced learners. Yet in a vocabulary learning experiment conducted with US university students of German, Stratton (2022) found that instruction in historical linguistics was helpful for learning German–English cognates that were not readily recognizable at first sight.

The prospective intralingual method with root families, though commonly used by expert language learners, appears to be unexplored in the scientific literature. The proposed textbook dictionary, in addition to being an immediate help for language students who want to expand their receptive lexical knowledge rapidly, could serve as a foundation for research into the effectiveness of explicit etymology – both interlingual and intralingual – as a pedagogical tool in second language learning.

References

August, G. (2009). *Wortfamilienwörterbuch der deutschen Gegenwartssprache* [Studienausgabe]. Niemeyer.

Stratton, J. (2022). Intentional and Incidental Vocabulary Learning: The Role of Historical Linguistics in the Second Language Classroom. *The Modern Language Journal,* 106(4), 837-857.

Tschirner, E. & Möhring, J. (2020). *Frequency dictionary of German: Core vocabulary for learners*. Routledge.

Winnerlöv, J. (2013). *Wie viele Wörter bekommt man beim Deutschlernen „geschenkt"? Der Kognatenstatus der häufigsten deutschen Wörter aus Sicht schwedischer LernerInnen auf Universitätsebene* [Magisteruppsats]. Stockholms universitet. urn:nbn:se:su:diva-91818

Contact information

Jonas Winnerlöv
European Parliament
jonas.winnerlov@europarl.europa.eu

Sylwia Wojciechowska

EXPLORING LARGE LANGUAGE MODELS IN WORD SENSE DISAMBIGUATION

A Study of Corpus Examples for Perception Verbs in Learners' Dictionaries

Keywords Large Language Models; word sense disambiguation; corpus examples; learners' dictionaries online; polysemy

Recent developments in Large Language Models (LLMs) have opened new possibilities for supporting lexicographic work. Studies suggest that LLMs can assist in tasks such as generating definitions, drafting example sentences, and assigning examples to dictionary senses (Lew, 2023; Jakubíček & Rundell, 2023). At the same time, the status of word senses remains one of the most debated issues in lexicography, with scholars highlighting the fluidity and context-dependence of meaning. Traditional lexicographic practice involves abstracting discrete senses from potentially unlimited numbers of contextualised corpus attestations. This process is particularly challenging in the case of highly frequent and polysemous lexical items, where meanings often overlap and form complex semantic networks. Dictionaries may adopt different approaches to sense division, "lumping" or "splitting" senses, which leads to considerable variation in sense inventories (Bond et al., 2024).

Nowadays, apart from assigning corpus-based examples to particular senses, online learners' dictionaries increasingly include large collections of automatically extracted corpus examples that are not explicitly linked to individual senses. While these collections provide valuable data, their pedagogical usefulness remains limited, as examples are not systematically organised. Recent advances in generative AI suggest that LLMs may help bridge this gap by analysing contextual meaning and aligning examples with sense definitions. Research indicates that such models are capable of handling metaphor, semantic relatedness, and graded meaning variation (Lin et al., 2024; Bond et al., 2024). Nevertheless, their performance in dealing with fine-grained sense distinctions and competing lexicographic models of word sense disambiguation remains underexplored.

The present study examines the potential of LLMs as tools for analysing sense granularity taking a cross-dictionary perspective. This paper investigates the ability of ChatGPT (GPT-5.5) to interpret corpus examples of English perception verbs and align them with the existing sense inventories. Both sense divisions and sections of unassigned corpus examples were drawn from two major online monolingual learners' dictionaries: *Cambridge Advanced Learner's Dictionary* and *Longman Dictionary of Contemporary English*. As the number of unassigned corpus examples for each examined entry differs between the dictionaries, the first twenty examples per verb were analysed. The dataset includes eleven high-frequency and highly polysemous perception verbs: *see*, *hear*, *touch*, *taste*, *smell*, *look*, *watch*, *observe*, *notice*, *listen* and *feel*. These verbs display systematic extensions from the physical domain into cognitive and evaluative domains, hence are suitable for examining sense granularity and semantic overlap. Given their systematic polysemy, the study examines how ChatGPT handles semantic shifts from physical to abstract and metaphorical meanings.

The following research questions are posed:

- To what extent does the success rate of ChatGPT in assigning corpus examples to dictionary senses depend on different lexicographic models of word sense disambiguation, lumping and splitting?
- Is ChatGPT able to identify cases of overlap and divergence of senses when assigning corpus examples?

For each target verb, sense definitions and accompanying examples of usage were compiled and incorporated into carefully structured prompts. These prompts instructed ChatGPT to match each unassigned corpus example to a numbered sense or to tag it as unmatched, as well as mark cases of overlapping or divergent senses in the dictionary entry on the basis of these examples. Two human experts independently evaluated the model's classifications and identified cases of agreement and disagreement.

The study points out candidate cases for sense lumping and splitting, as the model suggests close semantic relatedness between some senses that could be merged into one sense, or sense distinctions that are not reflected in the examined dictionaries. The findings demonstrate that LLMs can serve as useful tools for analysing sense granularity and assigning corpus examples to dictionary senses, provided that their output is subject to expert verification. As Bothma & Gouws (2025) emphasize, generative AI should be understood as an information tool rather than a replacement for expert judgement.

References

Cambridge advanced learner's dictionary (CALD). Retrieved May 25, 2026, from http://www.dictionary.cambridge.org

Longman dictionary of contemporary English (LDOCE). Retrieved May 25, 2026, from http://www.ldoceonline.com

Bond, F., Maziarz, M., Piotrowski, T., Rudnicka, E. (2024). Models of polysemy in two English dictionaries. *International Journal of Lexicography*, 37(2), 196-225.

Bothma, T. J. D. & Gouws, R. H. (2025). Generative Artificial Intelligence (GenAI) as information tool for lexicographic information needs. *Lexikos* 35(2), 276–295.

Jakubíček, M. & Rundell, M. (2023). The end of lexicography? Can ChatGPT outperform current tools for post-editing lexicography? In M. Medveď, M. Měchura, C. Tiberius, I. Kosem, J. Kallas, M. Jakubíček & S. Krek (Eds.), *Electronic lexicography in the 21st century: Invisible Lexicography. Proceedings of the eLex 2023 conference* (pp. 518-533). Lexical Computing.

Lew, R. (2023). ChatGPT as a COBUILD lexicographer. *Humanities and Social Sciences Communications*, 10(704). https://doi.org/10.1057/s41599-023-02119-6

Lin, Y., Liu, J., Gao,Y., Wang, A., Su, J. (2024). A dual-perspective metaphor detection framework using Large Language Models. arXiv. http://arxiv.org/abs/2412.17332

Contact information

Sylwia Wojciechowska
Adam Mickiewicz University, Poznań, Poland
Faculty of English
sylwiab@amu.edu.pl

Akerke Yessenali

PHRASEOLOGICAL THEORETICAL GAPS IN DIGITAL LEXICOGRAPHY

A Contrastive Study of English–German Phraseologisms with Gender Representations

Keywords Phraseology; digital lexicography; gender representation; contrastive analysis; English–German

Dictionaries are widely regarded as authoritative linguistic resources that play a central role in shaping and legitimising language use in society (Haß-Zumkehr 2001). Beyond their descriptive function, they act as social agents influencing public perceptions and norms (Gouws 2022) and must therefore be understood as socially embedded artefacts reflecting linguistic, cultural, and societal values (Müller-Spitzer 2023). This perspective is particularly relevant for gender representation, as lexicographic descriptions of personal referents such as woman and man have repeatedly been shown to reproduce asymmetrical and stereotypical patterns across languages (Hidalgo 2000; Kotthoff & Nübling 2018; Flood 2019; Diewald & Nübling 2022). Corpus-based studies, for instance, reveal systematic gendered asymmetries in adjectival co-occurrences, with men frequently characterised as *rich* or *strong* and women as *beautiful* or *pregnant* (Müller-Spitzer 2024). Public debates surrounding dictionary revisions, notably those concerning the Oxford Dictionaries (Flood 2019), further underscore the social impact of lexicographic decisions.

While previous research has primarily focused on individual lemmas and their typical co-occurrence patterns, the phraseological level remains comparatively underexplored in studies on gender representation. In particular, the categorisation and lexicographic treatment of idioms, semi-idioms, and collocations have received limited systematic attention (Burger et al. 2007; Burger 2015; Schafroth 2019). Existing analyses largely rely on condensed printed dictionaries, where phraseologisms may be typographically highlighted within single-word entries or occur unmarked in example sentences, complicating their identification and classification (Moon 2007; Steyer 2008; Hollós 2017). Even in comparable dictionaries, the selection, classification, and presentation of phraseological units frequently diverge (Moon 2007).

These challenges are further intensified in digital lexicography, where spatial constraints no longer apply and lexicographic practices vary considerably across platforms and languages (Gouws 2014; 2018; 2022; Klosa-Kückelhaus & Gouws 2015; Klosa-Kückelhaus 2024). Based on a manually compiled corpus drawn from selected German and English online dictionaries, comparable phraseologisms are frequently labelled, defined, positioned, or exemplified differently, or remain entirely unlabeled, directly affecting their identification in contrastive research. A case in point is the German semi-idiom *gefallenes Mädchen*, which appears as a separate hyperlinked entry under *Mehrwortausdrücke* in the *Digitales Wörterbuch der deutschen Sprache (DWDS)*, but occurs without any phraseological label in the example section of the *Duden Online-Wörterbuch* entry *Mädchen*. Its English equivalent *fallen woman* is treated equally inconsistently, appearing as a separate entry under *Other results* in *Oxford Learner's Dictionaries* and under *More results* in *Longman English Dictionary*, again without explicit phraseological labelling.

Against this background, the present research addresses a theoretical and methodological gap in digital lexicography by proposing a contrastive framework for the identification,

classification, and analysis of idioms, semi-idioms, and collocations with gender representations in German and English online dictionaries. The study is based on a manually curated contrastive lexicographic corpus compiled from selected online dictionaries and structured around a predefined set of key lemmas referring to socially defined personal roles, including general designations, family relations, romantic relationships, and social roles. The corpus-building process combines lemma-based extraction with manual phraseological annotation and cross-linguistic alignment. Phraseologisms were identified according to semantic opacity, degree of lexical fixation, and recurrent co-occurrence patterns, while the comparison across dictionaries focused on four analytical dimensions: phraseological labelling, positioning within entries, definitional treatment, and exemplification practices.

By foregrounding these methodological steps, the contribution outlines the compilation of a contrastive lexicographic corpus and demonstrates how inconsistencies in online dictionary structures require interpretative decisions during corpus annotation. It then examines how phraseological units are treated in major German platforms (e.g., *DWDS, Duden*) and English learner's dictionaries (e.g., *Oxford Learner's Dictionaries, Longman*), highlighting systematic differences in lexicographic practice.

The analysis reveals recurring cross-linguistic and platform-specific divergences in the treatment of phraseological units and gender representation. To address these inconsistencies, the paper proposes a flexible comparative annotation model that distinguishes between the lexicographic treatment of a unit (e.g., labelled or unlabelled, embedded or listed separately) and its linguistic properties as a phraseologism. This approach allows formally comparable units to be analysed despite divergent dictionary structures. The findings demonstrate that collocations pose particular methodological difficulties because their status as phraseologisms often remains implicit and platform-dependent. More broadly, the study argues that contrastive phraseological research in digital lexicography requires transparent annotation criteria and platform-sensitive analytical categories in order to account for the variability of online dictionary practices.

References

Burger, H. (2015). *Phraseologie. Eine Einführung am Beispiel des Deutschen*. Erich Schmidt Verlag.

Burger, H., Dobrovol'skij, D., Kühn, P. & Norrick, N. (2007). *Phraseologie – Phraseology. Ein internationales Handbuch der zeitgenössischen Forschung*. De Gruyter.

Diewald, G. & Nübling, D. (2022). Genus–Sexus–Gender – ein spannungs- und ertragreiches Themenfeld der Linguistik. In G. Diewald & D. Nübling (Eds.), *Genus – Sexus – Gender* (pp. 3–31). De Gruyter.

Flood, A. (2019). *Thousands demand Oxford dictionaries 'eliminate sexist definitions'*. The Guardian. Retrieved January 28, 2026, from https://www.theguardian.com/books/2019/sep/17/thousands-demand-oxford-dictionaries-eliminate-sexist-definitions

Gouws, R. H. (2014). Article Structures: Moving from Printed to e-Dictionaries. *Lexikos*, 24, 155–177.

Gouws, R. H. (2018). Internet Lexicography in the 21st Century. In S. Engelberg, H. Kämper & P. Storjohann (Eds.), *Wortschatz: Theorie, Empirie, Documentation* (pp. 215–236). De Gruyter.

Gouws, R. H. (2022). Dictionaries as Instruments of Exclusion and Inclusion: Some South African Dictionaries as Case in Point. *Lexicographica*, 38(1), 39–61.

Gouws, R. H. (2022). Extended Article Comments in Online Dictionaries. *Lexicography*, 9(2), 111–138.

Haß-Zumkehr, Ulrike. (2001). *Deutsche Wörterbücher – Brennpunkt von Sprach- und Kulturgeschichte*. De Gruyter.

Hidalgo Tenorio, E. (2000). Gender, sex and stereotyping in the Collins COBUILD English language dictionary. *Australian Journal of Linguistics*, 20(2), 211–230.

Hollós, Z. (2017). Wörterbuchkritik – Phraseologische Wörterbücher. In M. Bielińska & S. Schierholz (Eds.), *Wörterbuchkritik – Dictionary Criticism* (pp. 323–352). De Gruyter.

Klosa-Kückelhaus, A. (2024). *Internet Lexicography: An Introduction*. De Gruyter.

Klosa-Kückelhaus, A. & Gouws, R. H. (2015). Outer Features in E-Dictionaries. *Lexicographica*, 31, 142–172.

Kotthoff, H. & Nübling, D. (2018). *Genderlinguistik: Eine Einführung in Sprache, Gespräch und Geschlecht*. Narr.

Moon, R. (2007). Phraseology in general monolingual dictionaries. In H. Burger, D. Dobrovol'skij, P. Kühn & N. Norrick (Eds.), *Phraseologie – Phraseology. Ein internationales Handbuch der zeitgenössischen Forschung* (pp. 909–918). De Gruyter.

Müller-Spitzer, C. (2023). Gender stereotypes in dictionaries: the challenge of reconciling usage-based lexicography with the role of dictionaries as social agents. *Lexikos*, 33(2), 79–94.

Müller-Spitzer, C. (2024). Ein Mann ist reich, stark und bewaffnet und eine Frau schön, nackt und schwanger. Stereotype Zuschreibungen zu Mann und Frau in Pressetexten und ihr Einfluss auf korpusbasierte Wörterbücher des Deutschen. In A.-E. Ewels & D. Nübling (Eds.), *Geschlechterbewusste Sprache. Argumente, Positionen, Perspektiven* (pp. 269–294). Olms.

Schafroth, E. (2019). *Einführung in die Phraseologie-Tutorials. Phraseologie-Tutorials – ein digitales Lehrprojekt.* Retrieved January 28, 2026, from https://phraseologie.phil.hhu.de/intro/

Steyer, K. (2008). Kollokationen in deutschen Wörterbüchern und in der deutschen Wörterbuchforschung. *Lexicographica*, 24, 185–207.

Corpus sources:

DWDS: *Digitales Wörterbuch der deutschen Sprache.* Retrieved January 28, 2026, from https://www.dwds.de

Duden: *Duden Online-Wörterbuch.* Retrieved January 28, 2026, from https://www.duden.de

OLD: *Oxford Learner's Dictionaries.* Retrieved January 28, 2026, from https://www.oxfordlearnersdictionaries.com

LDOCE: *Longman Dictionary of Contemporary English Online.* Retrieved January 28, 2026, from https://www.ldoceonline.com

Contact information

Akerke Yessenali
ATILF (UMR 7118), CNRS – Université de Lorraine, Nancy, France
akerke.yessenali@univ-lorraine.fr

Manjola Zaçellari

THE MAKING OF MEANING

From the 1980 Communist Albanian Dictionary to the Great Dictionary of 2026

Keywords Communist; dictionary; Albanian

In the Albanian totalitarian political context, linguistic resources were considered an important tool for the installation of the guiding ideology and the materialist worldview. Dictionaries, being reference works, were very favourable and effective for normalizing any ideological perspective, even presenting them as objective linguistic truths. (Fairclough, 1992, p. 87). Our study aims at identifying the techniques of ideological manipulation imposed on the Albanian lexicographic practice of the communist period. We have taken into consideration The Dictionary of the Contemporary Albanian Language (1980) compiled during the communist period and compared it with The Great Dictionary of the Albanian Language (2026) published online in early 2026 with the aim of analysing how the ideological dimension is reflected in two different historical and methodological periods.

The research poses an explicit question regarding the influence of ideological constraints on lexicographic methodology, macrostructure, and microstructure in Dictionary of 1980, and their differences in relation to more descriptive and corpus-based approaches in Great Dictionary of the Albanian Language (2026). It further investigates which lexicographic tools, like lemma selection, definition strategies, illustrative examples, and usage labelling, carried ideological meaning in the period of communism. In its methodological design, the study is grounded in a qualitative and quantitative comparative analysis of the two dictionaries, addressing both macrostructural and microstructural levels of lexicographic analysis, including selection and organization of headwords, criteria of selection and exclusion, definitions, examples, and evaluative labels that reflect ideological interventions in lexicography. Particular attention is given to those examples and labels that express political and ideologically coloured meanings.

In compiling The Dictionary of 1980, Albanian linguists worked in a political environment, where Marxist-Leninist doctrine shaped both the language policies, and the scientific production. Consequently, this dictionary would reflect not only the codification of the linguistic norm, but also ideological interventions that affected both the macrostructure and the microstructure levels. Terminology from the political field, concepts related to religion, etc. were given a strong ideological content, while terms such as *socialism*, *internationalism* were often defined through positive explanations, typical of the language of propaganda.

The 1980 Dictionary was based entirely on Russian lexicographic practice, and the Soviet orientation even appears openly. (Lloshi, 2025, p.36). In fact, the explanation of the term *socialism* is a translation from the Dictionary of the Russian Language (vol. IV, 1961) and ends with the same words: 'the first phase of communism'. The meaning of the word *internationalism* is explained in the Dictionary of 1980 using an elevated style, with positive and evaluative elements in antagonism with the explanation for the word *nationalism*, a clear example of lexicographic megalomania according to communist lessons. In this case too, the meaning is partly taken from the Russian Dictionary definition of 1957-1961.

Meanwhile, the meaning of the word *capitalist* goes beyond the definition found in the 1957 Russian dictionary, because in the Albanian Dictionary of 1980 capitalists are described as

brutally exploiting wage workers and enriching themselves by appropriating surplus value. Likewise, the illustrative examples used the same language, often glorifying socialist morality and the narratives of the Party. Even for words like *Marxism* and *Leninism*, the illustration is given in the same way as in Russian dictionaries.

Meanwhile, at the beginning of 2026, the Great Dictionary of the Albanian Language was launched online. In contrast to the 1980s dictionary, it (the compilation of which included the author of this study) represents a modern lexicographic paradigm grounded in comprehensive documentation and descriptive methodology. This shift becomes more apparent when we analyse political or socially sensitive terms. For example, the word *capitalism* in the 2026 Dictionary is defined in completely neutral language. The same applies to terms such as *bourgeois, nationalist, identity, idea, moral*, etc.

These changes are evidence of the transition from ideologically prescriptive lexicography toward descriptive and corpus-based lexicographic methodology. At the same time, they reflect the processes of semantic neutralization, de-ideologization and reduction of the positive framing of the word meanings in contemporary Albanian lexicography.

References

Atkins, S. T., & Rundell, M. (2008). *The Oxford guide to practical lexicography*. Oxford University Press.

Béjoint, H. (2010). *The lexicography of English: From origins to present*. Oxford University Press.

Bergenholtz, H., & Tarp, S. (2002). Die moderne lexikografische Funktionslehre. *Lexicographica,* 18, 253-263. https://doi.org/10.1515/9783484604476.253

Fairclough, N. (1992). *Discourse and social change*. Polity Press.

Fjalori i madh i gjuhës shqipe. (2026). *Akademia e Shkencave e Shqipërisë*. Retrieved April 2, 2026, from https://fjalori.online

Institut russkogo iazyka Akademii nauk SSSR. (1957–1961). *Slovar' russkogo iazyka* [Dictionary of the Russian Language] (Vols. 1–4). Gosudarstvennoe izdatel'stvo inostrannykh i natsional'nykh slovarei.

Kostallari, A. (Kryered.), Thomai, J., Lloshi, Xh., et al. (1980). *Fjalor i gjuhës së sotme shqipe*. Akademia e Shkencave e Shqipërisë.

Lloshi, Xh. (2025). *Leksikografi: fjalorë të shqipes*. Albas.

Svensén, B. (2009*). A handbook of lexicography: The theory and practice of dictionary-making*. Cambridge University Press.

van Dijk, T. A. (1998). *Ideology: A multidisciplinary approach*. Sage.

Wiegand, H. E. (1989). Der gegenwärtige Status der Lexikographie und ihr Verhältnis zu anderen Disziplinen. *Lexicographica*, 5, 1–30.

Contact information

Manjola Zaçellari
Aleksander Moisiu University
manjolazacellari@uamd.edu.al

Tanara Zingano Kuhn, Vojtěch Kovář & Miloš Jakubíček

MAPPING SOCIOLINGUISTIC THEORY INTO PRACTICE

Headword Candidate Selection in the Dictionary of Pluricentric Portuguese

Keywords Headword candidates selection; pluricentric Portuguese; sociolinguistics

Presently in its preliminary phase at the Research Centre for General and Applied Linguistics at the University of Coimbra (CELGA-ILTEC), the dictionary of pluricentric Portuguese project seeks to develop an online dictionary that documents the usage of Portuguese across diverse global contexts. Considering the socio-historical complexities inherent to the Portuguese language area (Faraco, 2016; Correia, in print), this project's lexicographic work is grounded in a deeper understanding of language use and its social context, with sociolinguistic theory serving as a core guiding framework. The purpose of this paper is to present how theoretical insights from sociolinguistics translate into practical lexicographic decisions regarding headword candidate selection.

Portuguese is the official language of nine countries (Angola, Brazil, Cabo Verde, Guinea-Bissau, Equatorial Guinea, Mozambique, Portugal, Sao Tome and Principe, and Timor-Leste) and one territory (Macao), although its actual functional status varies significantly in these multilingual regions. Traditionally, Portuguese is considered to have two dominant varieties, Brazilian Portuguese and European Portuguese, with the latter being adopted as the norm in the other countries. The lexicography of the language reflects this bicentrism and dictionary production is largely centred in Brazil and Portugal (Kuhn & Correia, 2025). However, in countries where Portuguese was introduced as a result of colonisation (all countries except Portugal), research has revealed discrepancy between real used norms and the official discourse legislating the language, including teaching materials, reference instruments such as dictionaries, and official documents. In Brazil, for instance, the dominant exonormative view of the language has led to a significant gap between how language is actually used by formally educated people in contexts requiring more language monitoring, i.e., the cultivated standard, and its prescriptive use, or the ideal language standard (Faraco, 2008). Relatedly, studies on the real norms used in the other countries include the description of Mozambican Portuguese by Gonçalves P. (2010) and Firmino (2011), Angolan Portuguese by Adriano (2015) and Inverno (2009), and Sao Tomean Portuguese by Gonçalves R. (2016) and Bouchard (2017). Considering all this in this dictionary project requires, in addition to a robust foundational sociolinguistic element, the compilation of a corpus that reflects the real use of Portuguese, in a variety of contexts and in different territories. Since a corpus with these characteristics is not available, a decision was made to carry out a pilot-dictionary project, using a small corpus of tweets (Rodrigues Gomide, 2022). The purpose of the pilot is to try methodologies, evaluate results, and develop solutions to the problems found, in a smaller scale, alongside the compilation of a larger corpus and further construction of the theoretical framework. For that, we have adopted the Dictionary Express method (Baisa et al., 2019). The planned workflow consists of several steps of iterative collaboration between human lexicographers and automatically extracted data, with headword candidate selection as the first step. This involves headword annotation using a specially created interface, with different lexicographers assigning flags (e.g., not a lemma, non-standard, etc.) to the same set of candidates and final headword list definition resulting from general agreement among lexicographers. Differently from previous projects

for other languages (Baisa et al., 2019; Blahuš et al., 2023; Kovařík et al., 2024), in our project this method is used to make a dictionary that describes, at the moment, eight varieties of one language. One additional challenge stems from the sociolinguistic grounding of the project, by which traditional views on standard language are questioned. As a result, some changes have been made, such as the decision to not filter the list of automatically extracted candidate headwords for part of speech. This is because POS-taggers have been developed for Brazilian and European Portuguese, thus words from other varieties or words that are not considered to be standard might not be identified, leading to wrong POS-tagging which, in turn, would affect the frequency count. Another aspect of the headword candidate annotation process that is especially crucial to our project refers to the flag "non-standard". This type of decision-making requires a previous common-ground perspective on standardisation and standard language ideologies (McLelland, 2021) and how they affect the representation of language in the dictionary. This consideration becomes even more critical in the case of Portuguese, which is spoken across multilingual regions marked by complex socio-political-cultural contexts, particularly considering the impact of colonial history and the enduring legacy of colonial language policies. At the time of writing, this step of the workflow is still ongoing; however, further results are expected for the time of the conference. With the presentation of how the theoretical framing of this project maps into practical lexicographic work regarding headword candidate selection, we hope to share our experience with our peers, as well as contribute to questioning and changing long-established language ideologies and language attitudes prevailing in the Portuguese language area.

References

Adriano, P. S. (2015). *A crise normativa do português em Angola: Cliticização e regência verbal: Que atitude normativa para o professor e o revisor?* [The normative crisis of Portuguese in Angola: Cliticization and verbal regency: What normative attitude for teachers and proofreaders?] Maymba Editora.

Baisa, V., Blahuš, M., Cukr, M., Herman, O., Jakubíček, M., Kovář, V., Medveď, M., Měchura, M., Rychlý, P., & Suchomel, V. (2019). Automating dictionary production: A Tagalog-English-Korean dictionary from scratch. In I. Kosem, T. Zingano Kuhn, M. Correia, J. P. Ferreira, M. Jansen, I. Pereira, J. Kallas, M. Jakubíček, S. Krek & C. Tiberius (Eds.), *Electronic lexicography in the 21st century: Smart lexicography. Proceedings of the eLex 2019 conference* (pp. 805–818). Lexical Computing CZ s.r.o.

Blahuš, M., Cukr, M., Herman, O., Jakubíček, M., Kovář, V., Kraus, J., Medveď, M., & Ohlídalová, V. (2023). Rapid Ukrainian-English dictionary creation using post-edited corpus data. 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. 613–637). Lexical Computing CZ s.r.o.

Bouchard, M.-E. (2017). *Linguistic variation and change in the Portuguese of São Tomé* (PhD thesis). New York University.

Faraco, C. A. (2016). *História sociolinguística da língua portuguesa* [Sociolinguistic history of the Portuguese Language]. Parábola.

Correia, M. (pre-print). *O atual estatuto internacional do português– um percurso de cinquenta anos* [The Current International Status of Portuguese – A Fifty-Year Journey].

Faraco, C.A. (2008). *A Norma culta Brasileira: desatando alguns nós* [The Brazilian cultured norm: untying some knots]. Parábola.

Gonçalves, P. (2010). *A génese do português de Moçambique* [The genesis of Mozambican Portuguese]. Imprensa Nacional Casa da Moeda.

Gonçalves, R. (2016). *Construções ditransitivas no português de São Tomé* [Ditransitive constructions in Sao Tome Portuguese] (PhD thesis). University of Lisbon.

Inverno, L. (2009). *Contact-induced restructuring of Portuguese morphosyntax in interior Angola: Evidence from Dundo (Lunda Norte)* (PhD thesis). University of Coimbra.

Kovařík, F., Kovář, V., & Blahuš, M. (2024). On rapid annotation of Czech headwords: Analysing the first tasks of Czech Dictionary Express. In K. Š. Despot, A. Ostroški Anić & I. Brač (Eds.), *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress* (pp. 336–344). Institute for the Croatian Language.

Kuhn, T. Z., & Correia, M. (2025). *Portuguese language lexicography*. In Reference Module in Social Sciences. Elsevier.

McLelland, N. (2021). Language standards, standardisation and standard ideologies in multilingual contexts. *Journal of Multilingual and Multicultural Development*, *42*(2), 109–124. https://doi.org/10.1080/01434632.2019.1708918

Rodrigues Gomide, A. (2022). CPLP:tuítes – The pluricentric corpus of tweets in Portuguese language (1.0.0) [Data set]. *IV Encontro Internacional de Língua Portuguesa e Relações Lusófonas* (LUSOCONF 2022). Zenodo.

Contact information

Tanara Zingano Kuhn
University of Coimbra
tanarazingano@uc.pt

Vojtěch Kovář
Lexical Computing
vojtech.kovar@sketchengine.eu

Miloš Jakubíček
Lexical Computing
milos.jakubicek@sketchengine.eu

Acknowledgements

This work is funded by national funds through FCT – Foundation for Science and Technology, I.P., under project UID/04887/2025, and by international funds under the project https://doi.org/10.54499/UID/PRR2/04887/2025

Valeria Zotti & Rita Gramellini

FROM PLURILINGUAL LEXICOGRAPHY AND TERMINOGRAPHY TO GENERATIVE AI TRANSLATION

Potential, Limits, and the Enduring Role of Dictionaries

Keywords Plurilingual lexicography; terminography; Neural Machine Translation; generative artificial intelligence; specialized translation; artistic lexicon

Recent advances in generative artificial intelligence (AI) have profoundly reshaped the landscape of translation. Large Language Models (LLMs) such OpenAI's GPT series are now capable of producing fluent, coherent and context-aware translations that go beyond the word-for-word logic traditionally associated with earlier forms of Machine Translation (MT) systems (Kenny, 2022). Despite their apparent accuracy, however, these tools demonstrate uneven translation performance with texts belonging to different textual genres and degrees of specialization.

This paper investigates the potential and limits of current leading MT paradigms through the qualitative analysis of translations of texts situated along the continuum between general and specialized language (Cabré, 2007), with particular attention to semi-specialized discourse. The study draws on empirical data from a comparable corpus of travel literature and compares translations of French texts related to artistic and cultural heritage into English and Italian, produced by two Neural Machine Translation (NMT) systems (DeepL and Google Translate) and three generative AI models (ChatGPT-5.5, Claude Sonnet 4.6 and Mistral Medium 3.5).

The findings show that, while generative AI models often outperform traditional NMT systems in terms of fluency and overall textual coherence (Jiao et al., 2023; Hendy et al., 2023), they continue to display significant weaknesses in areas that are crucial for specialized and semi-specialized translation. These include inadequate handling of multi-word terms (e.g. *tableau d'autel*) and specialized lexical combinations (L'Homme, 1998, 2017) (e.g. *colonne ronde*), difficulties in semantic disambiguation (e.g. *tombeau*), and instances of so-called "hallucinations". Such limitations can be attributed to the probabilistic, data-driven nature of LLMs, which lack genuine semantic understanding and are trained predominantly on large but largely generic and unevenly distributed corpora (Bender et al., 2021; Ciotti, 2023; Raus, 2024; Tekwa, 2025).

Against this backdrop, the paper revisits a question often considered outdated in the "Machine Translation Era" (Asscher, 2025, p. 4): are traditional dictionaries still worth using? The analysis strongly supports an affirmative answer. General bilingual and monolingual dictionaries remain essential for managing polysemy, stabilizing semantic correspondences, and preventing uncontrolled lexical variation – tasks that generative models perform implicitly and inconsistently (Mattioda, 2024). Even more crucial is the role of specialized lexicographic and terminological resources, which provide explicit conceptual structuring, domain-specific distinctions, and historically grounded knowledge that AI systems are currently unable to guarantee.

Rather than being rendered obsolete by generative AI, monolingual and plurilingual dictionaries emerge as complementary resources supporting human pre- and post-editing processes,

guiding prompt design, and constituting high-quality supervised datasets for domain adaptation in neural and generative systems (Chu & Wang, 2018; Hansen, 2024). In this perspective, translation is increasingly configured as a hybrid activity in which human intelligence, lexicographic expertise, and AI interact complementarily rather than competitively.

The paper concludes that the future of translation does not lie in the replacement of dictionaries, but in their renewed centrality as linguistic, cognitive, and epistemological anchors within an evolving ecosystem of generative AI tools.

References

Asscher, O. (2025). *Machine translation and Translation Theory*. Routledge.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In M. C. Elish, W. Isaac & R. S. Zemel (Eds.), *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922

Cabré, M. T. (2007). Constituer un corpus de textes de spécialité. *Cahier du CIEL,* 20, 37–56.

Chu, C. & Wang, R. (2018). A Survey of Domain Adaptation for Neural Machine Translation. In E. M. Bender, L. Derczynski & P. Isabelle (Eds.), *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 1304–1319). Association for Computational Linguistics. https://aclanthology.org/C18-1111/

Ciotti, F. (2023). Minerva e il pappagallo. IA generativa e modelli linguistici nel laboratorio dell'umanista digitale. *Testo e Senso,* 26, 289–315. https://doi.org/10.58015/2036-2293/671

Hansen, D. (2024). *Évaluation experte d'un prototype d'aide à la traduction créative : la traduction littéraire automatique individualisée au regard de ses enjeux traductologiques, éthiques et sociétaux*. [Doctoral dissertation, Université de Liège - Université Grenoble Alpes]. ORBi-University of Liège. https://orbi.uliege.be/handle/2268/312631

Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H. & Awadalla HH. (2023). How good are GPT models at machine translation? A comprehensive evaluation. *arXiv*. https://doi.org/10.48550/arXiv.2302.09210

Jiao, W., Wang, W., Huang, J.-T., Wang, X., Shi, S., & Tu, Z. (2023). Is ChatGPT a Good Translator? A Preliminary Study. *arXiv*. https://doi.org/10.48550/arXiv.2301.08745

Kenny, D. (2022). Human and machine translation. In D. Kenny (Ed.), *Machine translation for everyone: Empowering users in the age of artificial intelligence* (pp. 1–24). Language Science Press.

L'Homme, M.-C. (1998). Caractérisation des combinaisons lexicales spécialisées par rapport aux collocations de langue générale. In T. Fontenelle, P. Hiligsmann, A. Michiels, A. Moulin & S. Theissen (Eds.), *Proceedings of the 8th EURALEX International Congress* (pp. 513–522). Université de Liège.

L'Homme, M.-C. (2017). Combinatoire spécialisée : Trois perspectives et des enseignements pour la terminologie. *TTR: Traduction, Terminologie, Rédaction*, 30(1–2), 215–241.

Mattioda, M. M. (2024). Intelligence artificielle et diversité linguistique : Quelle gestion équitable pour la garantie des droits linguistiques ? *Language Problems and Language Planning*, 40(2), 146–167. https://doi.org/10.1075/lplp.00118.mat

Raus, R. (2024). Deep learning e traduzione intralinguistica: Riformulare i testi della pubblica amministrazione in modo inclusivo. *Studi Italiani di Linguistica Teorica e Applicata*, 53(2), 350–367.

Tekwa, K. (2025). Artificial intelligence, corpora, and translation studies. In D. Li & J. Corbett (Eds.), *The Routledge handbook of corpus translation studies* (pp. 103–118). Routledge.

Contact information

Valeria Zotti
University of Bologna
valeria.zotti@unibo.it

Rita Gramellini
University of Bologna
rita.gramellini2@unibo.it

PART IV

SHORT ABSTRACTS

EURALEX 2026

Andrea Abel, Luca Ducceschi & Federico Villa

DETECTING SEMANTIC VARIATION BETWEEN GERMAN STANDARD VARIETIES THROUGH WORD EMBEDDING COMPARISON

Keywords Semantic variation; pluricentric German; lexicography; lexical semantics; South Tyrolean German

From a pluricentric perspective, German possesses several standard varieties, including the one used in South Tyrol (Italy). While automated lexicography has long focussed on differences of lexical forms across these varieties, meaning differences in formally identical words are more difficult to identify systematically. This study addresses such diatopic semantic variation by comparing usage patterns in corpora from Germany and South Tyrol. Using distributional evidence from word embeddings, it shows how corpus-based comparison can support lexicographic analysis by highlighting shared words whose meanings or typical contexts diverge across varieties. The focus lies on the linguistic interpretation of these patterns and on their relevance for describing variety-specific meanings. The results confirm existing lexicographic descriptions while also identifying additional candidates for future verification.

Contact information

Andrea Abel
Eurac Research
andrea.abel@eurac.edu

Luca Ducceschi
Eurac Research
luca.ducceschi@eurac.edu

Federico Villa
Independent researcher
federico.villa03@icatt.it

Luke Omoyemi Akinremi & Elena Martín-Cancela

BILINGUAL LEXICOGRAPHY UNDER RESOURCE ASYMMETRY

Designing Yoruba–German Dictionary Articles in Dialogue with Galician–German

Keywords Bilingual lexicography; Yoruba; Galician; resource asymmetry; dictionary article design; AI-assisted lexicography; cross-pair methodological dialogue

This paper addresses the challenge of designing bilingual dictionary articles under resource asymmetry by focusing on the Yoruba–German language pair and drawing methodological insights from the more developed Galician–German ecosystem. It proposes a resource-aware model of dictionary article composition that explicitly accounts for limited evidence by making uncertainty visible, using heterogeneous sources (corpora, pivot resources, expert input). Our approach involves dictionary-critical analyses of both Yoruba–German and Galician–German lexicographic resources and the transfer of microstructural design strategies from Galician to Yoruba. Case studies include Yoruba–German prototype entries (e.g. the verb *ṣe* and the nouns *ayé* and *akọ́nimọ̀ọ́gbá*) developed from these findings, incorporating sense divisions, translation equivalents, examples, grammatical/constructional notes and cross-references. The study also demonstrates a human-led use of AI tools to enrich example sentences and collocations in the prototypes. Main contributions are the concept of a resource-aware dictionary article and a cross-pair methodological dialogue, operationalised in prototype entries. These contributions provide practical guidelines for lexicographers of under-resourced languages and suggest scalable, AI-assisted workflows for bilingual lexicography under unequal resource conditions.

Contact information

Luke Omoyemi Akinremi
Universidade de Santiago de Compostela
lukeakinremi@gmail.com

Elena Martín-Cancela
Universidade da Coruña
elena.martin@udc.es

Iván Arias-Arias, María José Domínguez Vázquez,
María Teresa Sanmarco Bande & Carlos Valcárcel Riveiro

FROM CONTENT TO CODE AND BACK

Evaluating Agentic Coding in LLMs for Building R Shiny Dictionary Interfaces

Keywords LLMs; agentic coding; e-lexicography; R Shiny; user-friendliness

While recent scholarship highlights the generative capabilities of LLMs for drafting lexicographic entries, their "black box" nature poses challenges regarding training corpora, linguistic representativity, and semantic hallucination. We propose that this opacity is less problematic when LLMs are deployed not as linguistic experts, but as software engineers. By leveraging LLMs as agentic coding tools, lexicographers can bridge the gap between linguistic data and digital publication, democratizing access to custom lexicographic tools. This study investigates how LLMs conceptualize "user-friendliness" when tasked with designing lexicographic interfaces. We evaluate three leading models in agentic coding – Claude Opus 4.6, ChatGPT-5.5, and Gemini 3.1 Pro – using a zero-shot prompting strategy within a case study on noun valency. Each model generates an R Shiny application from curated CSV data, accompanied by explicit justifications for its design choices. The generated interfaces are assessed against criteria fundamental to e-lexicography: access structures, visual design, information architecture, and alignment with user-centred principles for language learners. Results reveal substantial differences across models in how they conceptualize and operationalize usability, with each model prioritizing fundamentally different design aspects. Crucially, not all models prove equally reliable as agentic coders for generating lexicographic interfaces that meet professional standards, highlighting the need for critical evaluation before adoption.

Contact information

Iván Arias-Arias
Universidade de Santiago de Compostela
ivanarias.arias@usc.gal

María José Domínguez Vázquez
Universidade de Santiago de Compostela
majo.dominguez@usc.es

María Teresa Sanmarco Bande
Universidade de Santiago de Compostela
mt.sanmarco@usc.es

Carlos Valcárcel Riveiro
Universidade de Vigo
carlos.valcarcel@uvigo.gal

Petra Bago & Ivana Eterović

COMPILING A 16TH-CENTURY CROATIAN DICTIONARY IN THE AGE OF AI

Challenges and Perspectives

Keywords Historical lexicography; Croatian language; sixteenth century; Gospels; multiscriptality; language variation; artificial intelligence

This paper presents the theoretical and methodological foundations for the development of an online dictionary of sixteenth-century Croatian, based on a corpus of ten sixteenth-century Gospel translations written in the Glagolitic, Cyrillic, and Latin scripts and representing four text types: missals, lectionaries, Bible translations, and postils. This undertaking constitutes one of the principal objectives of the project *CroBiLex: Language Variation in Croatian Books of the Early Modern Period* (*Based on the Gospels Corpus*) (Croatian Science Foundation, IP-2025-02-2567; 2026–2029), carried out at the Faculty of Humanities and Social Sciences, University of Zagreb. The paper outlines the proposed lexicographic architecture and discusses the challenges of processing a linguistically heterogeneous and multi-script corpus. Particular attention is devoted to the development of a Human-in-the-Loop workflow combining traditional philological methods with experimental applications of Large Language Models (LLMs) and LLM-assisted HTR/OCR technologies. The study examines the potential of LLMs in pre-lexicographic tasks: transcription, candidate lemma grouping, semantic clustering, and preliminary entry drafting, while also addressing the risks of historical inaccuracy and hallucinations in the processing of highly variable sixteenth-century Croatian texts. The proposed framework seeks to balance automation with scholarly accountability and to establish a sustainable digital model for historical Croatian lexicography.

Contact information

Petra Bago
University of Zagreb, Faculty of Humanities and Social Sciences, Department of Information and Communication Sciences
pbago@m.ffzg.hr

Ivana Eterović
University of Zagreb, Faculty of Humanities and Social Sciences, Department of Croatian Language and Literature
isankovi@m.ffzg.hr

Martina Bajčić & Ana Ostroški Anić

DEFINING LEGAL CONCEPTS IN THE AI AGE

An Exploratory Study into Influencer Practices

Keywords Legal concepts; definitions; user evaluation; AI; LLMs

Definitions play a crucial role in the acquisition, transfer, or translation of specialized knowledge. In the legal profession, precise definitions of complex concepts facilitate seamless knowledge management, application, and interpretation. However, despite growing reliance on artificial intelligence (AI) for this task, the real-world feasibility and user acceptability of AI-generated legal drafting remain largely unexamined. When addressing this gap, it is critical to consider that the use of authoritative legal terminology is central to the validity of any legal definition. Indeed, precise terms facilitate the application and interpretation of core concepts. However, new legal domains lack stable legal terms and conceptually demarcated concepts. Resolving the inherent tension between linguistic evolution and the need for legal certainty make the legal domain a unique use case that necessitates domain-specific task adaptation for general-purpose large language models (LLMs). With that in mind, this paper evaluates the potential of LLMs to generate accurate legal definitions through an empirical study, testing a generated dataset of key influencer-related concepts across domain user groups to glean insights into their evaluations of the generated definitions. The results underscore the critical importance of data quality and precise prompt engineering when leveraging LLMs for complex, domain-specific tasks.

Contact information

Martina Bajčić
University of Rijeka, Faculty of Law
martina.bajcic@pravri.uniri.hr

Ana Ostroški Anić
Institute for the Croatian Language
aostrosk@ihjj.hr

Elena Berthemet

DESIGNING AN IDIOMATIC DICTIONARY

Keywords Equivalence; microstructure; motivation; production; register; semantic relations; understanding; variation

This paper presents the principles and structure of *Colidioms*, an online idiomatic dictionary designed to provide a fine-grained description of idioms for their full understanding and production across typologically different language families, including French, Japanese, Korean, and Chinese. Grounded in cognitive and corpus linguistics and in recent work in phraseology and phraseography, the tool offers a unified model for both source and target languages, describing each idiom on at least three levels: formal structure, meaning(s), and use. The methodology combines corpus-based analysis, introspection, and linguistic experiments with native and non-native speakers. The paper focuses on the microstructure of entries, namely the canonical form and combinatorial properties, corpus-based definition(s), motivation, lexical, semantic, and syntactic co-occurrence constraints, register labels, variants, semantic links to related idioms (synonyms, equivalents, conversives, antonyms, false friends, hypernyms), notions, and pedagogical examples. The resource is both source- and target-oriented and enables multidirectional search for equivalents. We argue that *Colidioms* serves both practical aims (decoding and encoding idioms) and theoretical ones (contributing to the description of phraseological systems in typologically different languages and clarifying the theoretical status of idioms).

Contact information

Elena Berthemet
Centre de Linguistique en Sorbonne
elenaberthemet@free.fr

Harald Bichlmeier

VOL. IX OF THE *ETYMOLOGICAL DICTIONARY OF OLD HIGH GERMAN* [*ETYMOLOGISCHES WÖRTERBUCH DES ALTHOCHDEUTSCHEN (EWA)*]

History Structure, Contents

Keywords Etymological dictionary; Old High German; etymology; Germanic; Proto-Indo-European

The paper presents vol. IX of the *Etymological Dictionary of Old High German* [*Etymologisches Wörterbuch des Althochdeutschen*], the most encompassing etymological dictionary of any (Old) Germanic language. The volume will have roughly 800 pages, contain a total of 2,950 entries, of which 234 are main entries (the number is only about half the number of the main entries of vol. VIII) coming with a presentation of the full etymology and history of the word. Of these 234 main entries over 40 turn out to be loanwords, mostly attained from Medieval Latin (though originally sometimes from elsewhere). About 25 % of all entries (about 60 % of the main entries and almost 22 % of the fillers) are continued in modern Standard German or in German dialects. With these percentages the volume takes a middle position among the last volumes

Contact information

Harald Bichlmeier
Sächsische Akademie der Wissenschaften zu Leipzig
harald.bichlmeier@uni-jena.de

Riccardo Billero

LEXICOGRAPHY IN THE AGE OF AI

Training Systems with Human Expertise in Art Terminology

Keywords Specialized lexicography; artificial intelligence; human expertise; art terminology; human-AI collaboration; cultural heritage; multilingual lexicography; LBC project

The emergence of artificial intelligence represents the latest in a series of technological challenges that have periodically reshaped the relationship between human expertise and technological tools. This paper takes as its point of departure the conviction that AI cannot exist, function, or evolve independently of human intelligence; lexicography, as a discipline situated at the intersection of linguistic knowledge, cultural competence, and technological practice, is particularly well placed to explore how human expertise can guide AI systems rather than be displaced by them. Drawing on the resources of the *Lessico dei Beni Culturali* (LBC) project, a multilingual lexicographic initiative at the University of Florence covering Italian cultural heritage terminology across Italian, and seven further languages, the paper investigates how human-curated reference materials and expert-designed prompts can systematically improve AI performance on culturally embedded art terms. Three representative terms (*cartone*, *grazia*, and *maniera*) were tested across three conditions, namely baseline, expert-prompted, and LBC-constrained, using three major AI systems (Claude Sonnet 4.6, Gemini 3.5 Flash, and GPT-5). The results confirm a consistent gradient: corpus evidence not only improves output quality but, in some cases, reveals lexicographic realities that expert prompting actively conceals, most strikingly the absence of a stable specialised equivalent for *maniera* in Spanish, which the LBC corpus exposes as a structural lexical gap rather than a straightforward translational choice. The study argues that the value of expert lexicographic knowledge does not diminish in AI-assisted workflows but becomes more concentrated, shifting from the direct production of entries to the design of the knowledge infrastructure within which AI systems can operate effectively.

Contact information

Riccardo Billero
Free University of Bozen-Bolzano
rbillero@unibz.it

Ben Black Bear, Jr., Richard Two Dogs & Jan Ullrich

BUILDING THE LARGEST DICTIONARY FOR AN INDIGENOUS LANGUAGE OF NORTH AMERICA

Keywords Endangered language; bilingual dictionary; corpus; polysynthetic language; print and electronic dictionary

This study offers a brief overview of the work that led to the creation of what is currently the largest dictionary for an indigenous language of North America. It highlights our efforts to create a text corpus-based dictionary for Lakota, a highly polysynthetic and endangered language. We discuss how our team handled various issues stemming from limited funding and working with an under-resourced and under-documented language. Despite the various limitations, the New Lakota Dictionary now contains 47,000 entries and is available both in print and as an electronic application with features that match print and electronic dictionaries on modern languages, such as lemmatization, valency numbers, audio component of entries and example sentences, advanced corpus searches, inflection charts, and both monolingual and bilingual definitions, among others. Our work resulted in a highly comprehensive and reliable description of the Lakota lexicon, as well as in refining the linguistic analysis of Lakota lexical categories and their functions in Lakota morphosyntax.

Contact information

Ben Black Bear, Jr.
Lakota Language Consortium
benblkbr@lakhota.org

Richard Two Dogs
Lakota Language Consortium
r.twodogs@gmail.com

Jan Ullrich
Lakota Language Consortium
jfu@lakhota.org

Marek Blahuš, Miloš Jakubíček, Katarína Petreková & Yingtian Wu

AUTOMATIC DRAFTING OF TERMINOLOGICAL DICTIONARIES WITH ONE-CLICK DICTIONARY FOR 34 LANGUAGES

Keywords Automatic dictionary drafting; terminology; Sketch Engine; lexonomy; One-Click Dictionary

We present an evaluation of a system for automatic drafting of terminological dictionaries using the One-Click Dictionary feature of Sketch Engine. We bootstrapped five small corpora in Arabic, Czech, English, German and Spanish and evaluated the performance of the methodology on these five corpora yielding five terminological dictionaries for five different terminological domains; the system currently supports 34 languages in total.

Contact information

Marek Blahuš
Lexical Computing, Brno, Czech Republic
marek.blahus@sketchengine.co.uk

Miloš Jakubíček
Lexical Computing, Brno, Czech Republic
milos.jakubicek@sketchengine.eu

Katarína Petreková
Lexical Computing, Brno, Czech Republic Faculty of Arts, Masaryk University, Brno, Czech Republic
katarina.petrekova@sketchengine.eu

Yingtian Wu
Lexical Computing, Brno, Czech Republic
wyt15858810001@163.com

Igor Boguslavsky, Vyacheslav Dikonov, Evgeniya Inshakova, Alexandre Lazursky, Svetlana Timoshenko & Tatiana Frolova

LEXICOGRAPHIC RESOURCES FOR INFERENCE-BASED SEMANTIC TEXT PROCESSING SYSTEM

Keywords Morphological dictionary; combinatorial dictionary; ontology; repository of individuals; inference; digital linguistic resources

This paper describes the lexicographic resources developed for the ETAP-4 text analysis and generation system, with particular emphasis on its semantic module, SemETAP. In our approach, semantic analysis is viewed not merely as the construction of a semantic representation, but as the derivation of inferences licensed by linguistic and background knowledge. The semantic component operates in two stages. First, it constructs a Basic Semantic Structure (BSemS), which represents the meaning explicitly conveyed by the text. This representation is then enriched through ontological knowledge and inference rules, yielding an Extended Semantic Structure (EnSemS) that incorporates both explicit and implicit information. To support this process, ETAP-4 relies on a set of large-scale, knowledge-rich lexicographic resources characterized by a high degree of reusability. Linguistic knowledge is distributed across a morphological dictionary, a combinatorial dictionary, and rule-based grammars, whereas background knowledge is represented in an ontology, a repository of individuals, and a collection of inference rules. The paper describes the information encoded in these resources, shows formal languages used to represent it, and the interaction between linguistic and extra-linguistic knowledge during semantic analysis.

Contact information

Igor Boguslavsky
Universidad Politécnica de Madrid
igor.m.boguslavsky@gmail.com

Vyacheslav Dikonov
Institute for information transmission problems of the Russian Academy of Sciences (Kharkevich Institute)
dikonov@iitp.ru

Evgeniya Inshakova
Institute for information transmission problems of the Russian Academy of Sciences (Kharkevich Institute)
e.s.inshakova@gmail.com

Alexandre Lazursky
Institute for information transmission problems of the Russian Academy of Sciences (Kharkevich Institute)
lazursky@mail.ru

Svetlana Timoshenko
Institute for information transmission problems of the Russian Academy of Sciences (Kharkevich Institute)
nyrestein@gmail.com

Tatiana Frolova
Institute for information transmission problems of the Russian Academy of Sciences (Kharkevich Institute)
tfrolova@gmail.com

Ivana Brač

SEMANTIC CLASS ASSIGNMENT IN A CROATIAN VERB LEXICON

Challenges of Applying VerbNet's Classes

Keywords Verb lexicon; semantic classes; VerbNet; WordNet; Croatian

This paper examines the challenges of assigning VerbNet semantic classes to Croatian verb senses, based on data from the Croatian verb lexicon Verbion. The lexicon provides a multi-level description of the 500 most frequent verbs in Croatian, integrating lexical, semantic, syntactic, and morphological information. Verb senses are classified into semantic classes according to WordNet and VerbNet, using English equivalents as an intermediary. Since VerbNet is based on Levin's syntactic-semantic classification of English verbs, applying these classes to Croatian presents certain challenges. The analysis identifies five types of cross-linguistic mismatch. First, differences in lexicalization patterns reflect language-specific conventions governing what semantic components are encoded within the verb itself versus what is expressed syntactically. Second, Croatian reflexive morphology frequently introduces semantic distinctions that have no direct counterpart in VerbNet classes. Third, alternation-based classification cannot always be applied, as such alternations are either restricted or encoded differently in Croatian. Fourth, some VerbNet classes do not adequately capture the syntactic and semantic properties of Croatian verbs. Finally, certain high-frequency verbs are absent from the VerbNet inventory altogether. These findings suggest that VerbNet requires systematic adaptation for Croatian. The study contributes to the development of a Croatian-specific extension while preserving its cross-linguistic comparative potential.

Contact information

Ivana Brač
Institute for the Croatian Language
ibrac@ihjj.hr

Katja Brankačkec & Barbora Štěpánková

MULTIFUNCTIONAL FUNCTION WORDS IN THE UPPER SORBIAN–CZECH DICTIONARY MUDRA 2.0

Keywords Upper Sorbian; function words; bilingual dictionary; minority language; grammaticalization

This paper focuses on the lexicographic treatment of multifunctional function words in the emerging Upper Sorbian–Czech dictionary Mudra 2.0. As Upper Sorbian lexicography remains strongly dependent on German and lacks a monolingual explanatory dictionary, the compilation of bilingual Upper Sorbian-"non-German" dictionaries poses specific methodological challenges. These are particularly evident in the treatment of expressions with predominantly grammatical or pragmatic functions, such as conjunctions, modal words, and intensifiers. These expressions are often characterised by polysemy and multifunctionality resulting from grammaticalization processes, and their meanings are highly context-dependent. The treatment of these expressions in monolingual dictionaries has been inconsistent so far, whereas in bilingual dictionaries it is typically limited to providing translation equivalents. In this paper, we discuss the various possibilities and limitations of lexicographic descriptions of function words in the context of the minority language. Through a case study of the Upper Sorbian expressions *docyła* ('completely, entirely') and *cyle* ('completely, quite'), we analyse their semantic and pragmatic properties in comparison to their equivalents in Czech and German. Based on this analysis, we present the lexicographic approach adopted in the dictionary, including sample entries.

Contact information

Katja Brankačkec
Slovanský Ústav Akademie věd ČR
katja.brankackec@slu.cas.cz

Barbora Štěpánková
Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University
stepankova@ufal.mff.cuni.cz

Cormac Breathnach & Pádraig Ó Mianáin

REVERSING THE NEW ENGLISH–IRISH DICTIONARY DATABASE TO CREATE ENTRY FRAMEWORKS FOR IRISH–ENGLISH AND MONOLINGUAL IRISH DICTIONARIES

Keywords Monolingual dictionary; bilingual dictionary; reversal; Irish-language lexicography

The first major monolingual Irish dictionary, *An Foclóir Nua Gaeilge* [The New Irish Dictionary] was launched in December 2025, when an initial tranche of 40,000 senses was published on focloir.ie. The dictionary will be added to incrementally in 2026 and 2027 to bring it to its content target of 80,000 senses. Along with the monolingual dictionary, a new bilingual Irish-English dictionary is also being produced and will be fully published in 2027. The entry frameworks for the new project were created by reversing the *New English-Irish Dictionary* (NEID) (2013–2017) database. This paper explores the decision to reverse the NEID database, the challenges involved in specifying the task, the resulting Irish-language entry frameworks and their use, the unintended consequences and related challenges, and the solutions found to those challenges.

Contact information

Cormac Breathnach
Foras na Gaeilge
cbreathnach@forasnagaeilge.ie

Pádraig Ó Mianáin
Foras na Gaeilge
pomianain@forasnagaeilge.ie

Rukiye Smilla Burkart & Susanne Kabatnik

WIE DEFINIERT MAN DAS UNBEKANNTE?

Prompt Engineering als Methode der KI-gestützten Bedeutungsdefinition im Pandemictionary

Keywords Prompt engineering; digital lexicography; historical pandemic vocabulary; artificial intelligence; definition of meaning; Pandemictionary

The formulation of definitions is one of the core tasks of lexicography but has traditionally involved a great deal of manual work. With the availability of large language models, new opportunities have opened for automating lexicographical workflows. Although AI-assisted definition of meanings has already been thoroughly tested for 'well-known' vocabulary, this has not yet been the case for 'lesser-known', historical vocabulary, as such terms are scarcely represented in the training data of large language models. Hence, this paper investigates the extent to which targeted prompt engineering can improve the quality of AI-generated definitions for historical pandemic vocabulary. The data used in this study is drawn from the cholera corpus developed as part of the *Pandemictionary* project. Using eight headwords of varying frequency and semantic complexity, three prompt configurations are compared, with a particular focus on analysing the role of corpus evidence in the generation of lexicographical definitions.

The results show that AI systems are fundamentally capable of generating formally correct and lexicographically structured definitions of meaning. Whilst prompts without evidence often draw on general model knowledge and can therefore produce anachronistic or hallucinated content, evidence-based prompts lead to definitions that are more historically accurate and more firmly grounded in the corpus. The incorporation of corpus evidence proves to be a crucial quality factor, particularly in the case of rare or poorly documented lexemes. At the same time, challenges remain, such as those relating to hallucinations, polysemy and the reduction of definitions to their essential elements. This study highlights the potential of prompt engineering as a methodological tool in digital lexicography, whilst also demonstrating the continuing central role of human expertise in quality assurance.

Contact information

Rukiye Smilla Burkart
Trier University
burkart@uni-trier.de

Susanne Kabatnik
Trier University
kabatnik@uni-trier.de

Valeria Caruso & Geraint Paul Rees

FROM LEXICAL THEORY TO AI TRANSLATION

Procomplement Verbs and ChatGPT

Keywords Procomplement verbs; phraseology; Generative Lexicon; bilingual AI-assisted lexicography; AI-assisted translation; ChatGPT; LLMs

Procomplement verbs (PCVs) are a highly frequent yet understudied class of semantically idiosyncratic verb–clitic constructions found across Romance languages. Despite their relevance for everyday communication, they remain only marginally represented in grammars and dictionaries. Building on recent theoretical work on Italian PCVs and drawing on Generative Lexicon theory, this study investigates the relationship between the syntactic–semantic properties of PCVs and their treatment by large language models. A dataset of 160 highly frequent Italian PCVs, primarily extracted from GRADIT and comprising 256 usage examples, was submitted to ChatGPT 5.2 for translation from Italian into English. The analysis focuses on the model's ability to account for the semantic, pragmatic, and phraseological complexity of these constructions. The results show that translation quality is influenced less by the specific subclass of PCVs than by transversal properties such as connotation and the pragmatic functions encoded by individual utterances. While ChatGPT generally succeeds in translating strongly conventionalised and highly idiomatic expressions, it frequently encounters difficulties with polysemous PCVs, context-sensitive pragmatic meanings, and morphologically complex verb–clitic combinations. The findings highlight the value of a fine-grained syntactic–semantic classification of PCVs for bilingual lexicography and suggest potential applications for the improvement of AI-assisted translation systems.

Contact information

Valeria Caruso
Università degli Studi di Napoli "l'Orientale"
vcaruso@unior.it

Geraint Paul Rees
Universitat Pompeu Fabra
geraintpaul.rees@upf.edu

Lian Chen, Damien Nouvel, Huy-Linh Dao &
Alexander Delaporte

FROM CORPUS TO DICTIONARY

OntoLex-Lemon-Based Modelling of Neology in Chinese–French and Vietnamese–French Computational Lexicography

Keywords Neology; computational lexicography; OntoLex-Lemon; Multilingual Lexical Resources; semantic web; electronic dictionaries; Chinese; Vietnamese

This paper presents NeoLex, a multilingual electronic lexical resource dedicated to contemporary Chinese–French and Vietnamese–French neology. Based on diachronic media corpora (2015–2025) and pre-2015 reference lexicons, the project combines corpus linguistics, Natural Language Processing, and computational lexicography to detect emerging lexical units. Candidate neologisms were extracted using Out-of-Vocabulary detection, Vietnamese n-gram extraction, and a character-based method for Chinese, then manually validated according to lexicographic criteria. The validated entries were modelled using OntoLex-Lemon, SKOS, and vartrans to represent lexical information, semantic domains, multilingual correspondences, and diachronic metadata. A Streamlit-based prototype was developed to explore and visualize the resulting resources. NeoLex demonstrates how corpus-based neology detection and ontology-driven lexicography can contribute to interoperable multilingual electronic dictionaries.

Contact information

Lian Chen
LLL, University of Orleans
loselychen@gmail.com

Damien Nouvel
ERTILM
damien.nouvel@inalco.fr

Huy-Linh Dao
CRLAO-CNRS
huy-linh.dao@inalco.fr

Alexander Delaporte
CNRS-CRLAO
alexander.delaporte@ehess.fr

Luisa Cimander

ENGLISH ADJECTIVES IN GERMAN

Processes of Integration since the 17th Century and Lexicographic Implications

Keywords English loan adjectives; German adjectives; English loanwords in German; linguistic integration of loanwords; diachronic integration processes; lexicographic registration of loanwords

This paper presents the current state of a PhD project on the integration of English adjectives into German. While English nominal and verbal loanwords have undergone thorough investigation, adjectives were often only covered marginally. Some researchers have approached English adjectives more systematically and have identified integration processes and problems, but they mostly examined small data sets or specific genres (e.g. newspapers) at a certain point in time. For this PhD project, adjectives were collected from 1650 up to 2025 to create a substantial, diachronic basis for analysis. This large time span was further divided into three time periods, based on major turning points in 1870 and 1945, to render diachronic comparisons possible. The sources were scientific works as well as various types of dictionaries (e.g. general, foreign words, Germanising). First results show an increase of adjectives between the time periods, with a total of roughly 1000 adjectives in the collection. Currently, the adjectives are being categorised according to their morphological, phonological and orthographical properties, with the assumption being that adjectives with the same properties behave in the same way in regards to inflection and syntactic usage. These groups will later be verified through further (corpus) analyses.

Contact information

Luisa Cimander
Leibniz-Institut für Deutsche Sprache, Mannheim
cimander@ids-mannheim.de

Li Cui, Seol Namkung, Jun Lee, Hae-Yun Jung & Kilim Nam

WHAT LLMS LEARN AND WHAT THEY MISS

Frequency Effects in the Automated Description of Korean Verbal Inflection

Keywords LLM-assisted lexicography; low-resource languages; Korean verbal inflection; lexicographic automation; long-tail problem; low-frequency words; agglutinative language

The present study aims to empirically examine the possibilities and limitations of LLMs as a lexicographic tool for Korean, by automatically generating inflected forms of Korean verb and adjective headwords using Large Language Models and contrasting the results with corpus frequency data. Focusing on the description of inflected forms of Korean verbs and adjectives, the study analyses in detail the extent to which LLMs reproduce corpus-based frequency rankings and the accuracy of inflected form generation, divided into the categories of regular, irregular, and defective predicates. The results show that LLMs consistently produce various errors in low-frequency inflected forms and reveal clear limitations in precisely reproducing frequency information. This empirical examination points to the possibilities and risks inherent in LLMs as lexicographic assistance tools from multiple angles. While LLMs hold meaningful value as an assistive means for drafting dictionary descriptions, lexicographic reliability can only be guaranteed when this is complemented by systematic corpus-based verification and the distinctive expertise of lexicographers. These findings reaffirm the fundamental role that lexicographers must play in the automation of dictionary compilation.

Contact information

Li Cui
Yonsei University
choiri20@yonsei.ac.kr

Seol Namkung
Yonsei University
sul2001sul@naver.com

Jun Lee
Yonsei University
leejun0624@gmail.com

Hae-Yun Jung
Kyungpook National University
haeyun.jung.22@gmail.com

Kilim Nam
Yonsei University
nki@yonsei.ac.kr

Flávia de Oliveira Maia-Pires & Silvio Licurgo Pires

USER-ORIENTED SPECIALIZED DIGITAL LEXICOGRAPHY AND TERMINOLOGICAL ACCESSIBILITY

The LexiC Platform Applied to the Glossary of Brazilian Migration Legislation

Keywords Specialized lexicography; digital lexicography; user orientation; terminological accessibility; terminography; migration legislation

The technical and legal language of Brazilian migration legislation constitutes a significant barrier to information access for migrants and refugees, particularly those in situations of social, linguistic, and legal vulnerability. This paper presents the LexiC Platform, a web-based computational infrastructure for the collaborative management of terminological data, and demonstrates its application in the development of the Computerized Terminological Glossary of Brazilian Migration Legislation. The proposal articulates the communicative approach to Terminology (Cabré, 1999, 2005), functional user-oriented lexicography (Tarp, 2008, 2012), and Corpus Linguistics methodologies (Biber, Conrad & Reppen, 1998). The Platform organizes terminographic work into interdependent workflows corresponding to the stages of structured data entry, methodological and conceptual validation, distributed collaborative revision, and automated generation of lexicographic products, operationalized through computerized terminographic records with more than 60 standardized fields. The specialized corpus, comprising approximately 150,000 words, covers laws and decrees published between 2017 and 2025, with around 270 terms identified at different stages of description. The results indicate that the combination of terminographic rigor, collaborative digital infrastructure, and user orientation produces terminological resources that are both accessible and socially relevant.

References

Biber, D., Conrad, S. & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge University Press.

Cabré, M. T. (1999). *La terminología: representación y comunicación*. Institut Universitari de Lingüística Aplicada.

Cabré, M. T. (2005). *Terminology: Theory, methods and applications*. John Benjamins.

Tarp, S. (2008). *Lexicography in the borderland between knowledge and information*. Max Niemeyer.

Tarp, S. (2012). Lexicography and user needs. *Lexikos*, 22, 1–18.

Contact information

Flávia de Oliveira Maia-Pires
University of Brasília (UnB)
projetosmaiapires@gmail.com

Silvio Licurgo Pires
University Center of Brasília (UniCEUB)
licurgoslp@gmail.com

Gilles-Maurice de Schryver

CHATBOT LEXICOGRAPHY

Are GenAI LLMs Simply 'Dictionaries' in their Own Right and 'Lexicographers' Plain and Simple?

Keywords Artificial intelligence (AI); generative AI (GenAI); ChatGPT; Gemini; DeepSeek; Copilot; Claude; bibliometrics; lexicography

At the time of the Euralex 2026 conference it will be nearly four years since the start of the GenAI revolution. The question that needs to be asked, and which will be answered, is thus: Is any of the past work in (academic) lexicography still of any relevance, or are LLM chatbots simply 'dictionaries' in their own right and 'lexicographers' plain and simple? The answer is a shocking one.

Contact information

Gilles-Maurice de Schryver
Ghent University, Department of Languages and Cultures: Africa
gillesmaurice.deschryver@UGent.be

Gilles-Maurice de Schryver

THE NINTH ART IN LEXICOGRAPHY

The Power of Comic Strips

Keywords Ninth Art; comic strips; intersemiotic translation; lexicography; bibliometrics; artificial intelligence (AI); generative AI (GenAI); large language model (LLM); Gemini; ChatGPT; Academia.edu

This paper presents the background to a book on the visual history of academic lexicography, being a sequence of hundreds of one-page comic strips, created through conversations with LLM chatbots and GenAI image generators, each page representing a major publication in the field. Of course I knew that this submission would be a gamble: Would a comic version of a conference paper be taken seriously? Well, one extremely positive adjudicator gave it a 'Strong Accept', and a more critical one gave it an 'Accept'. Adjudicator 1 summarised it head-on, writing that it is "an extremely original contribution to the field, as it addresses the impact and outreach of scientific research both within and beyond academia", while Adjudicator 2 pointed out that it is "untraditional but entertaining, and comic strips could very well be part of lexicography." In what follows, my original extended abstract is presented first, after which I interact with the feedback left by Adjudicators 1 and 2. No doubt, and again, a very uncommon approach, but times have changed in the era of generative AI.

Contact information

Gilles-Maurice de Schryver
Department of Languages and Cultures: Africa, Ghent University
gillesmaurice.deschryver@UGent.be

Ida Dringó-Horváth & Katalin P. Márkus

DICTIONARY USE IN A TECHNOLOGY-DRIVEN ERA

Longitudinal Evidence from EFL and GFL Graduates

Keywords Dictionary use; dictionary skills; dictionary didactics; longitudinal research; AI-based tools

This paper presents findings from the third phase of a longitudinal repeated cross-sectional study examining dictionary-use patterns among graduates of English as a Foreign Language (EFL) and German as a Foreign Language (GFL) programmes in Hungary. Drawing on questionnaire data collected in 2020, 2023, and 2026, the study investigates how dictionary use is reshaped in an increasingly digital and AI-mediated reference environment, with particular attention to generative AI-based tools. The results indicate both continuity and change in reference practices. Online dictionaries and search engines remain widely used, but their dominant position is increasingly complemented by machine translation systems and generative AI tools. Print dictionaries continue to be widely owned, yet their frequency of use shows a gradual decline, pointing to a growing gap between ownership and actual use. The 2026 data provide early evidence that chatbots have become part of graduates' language-support repertoires, functioning as multifunctional tools that combine dictionary-like, translation, and text-support functions. The findings suggest a redistribution of reference practices within a hybrid digital environment and highlight the need to reconsider dictionary didactics in light of learners' expanding engagement with AI-based language technologies.

Contact information

Ida Dringó-Horváth
Karoli Gaspar University of the Reformed Church in Hungary
dringo.horvath.ida@kre.hu

Katalin P. Márkus
Karoli Gaspar University of the Reformed Church in Hungary
p.markus.kata@kre.hu

Sophia Fiedler

TOWARDS AN INTERACTIONAL LEXICON OF SPOKEN GERMAN

A Functional, Audio-Based Lexicographic Resource

Keywords Online lexicography; interactional linguistics; spoken language; dual access structure; German

This paper introduces *The interactional lexicon*, an online lexicographic resource for lexical units in German talk-in-interaction. The resource's core contribution is a function-first entry model that complements form-based access. Grounded in Interactional Linguistics, all entries are derived from empirical, peer-reviewed studies and document units that are frequent or distinctively used in spoken German, with transcripts from naturally occurring conversations and access to audio/video recordings (FOLK-corpus, IDS Mannheim). To accommodate the organizational specificities of interaction, we propose a compact entry schema: headword and function; position in the turn; sequential position; a corpus-linked example and an analysis thereof. Further categories for prosody and activity framework are optional. A dual access structure supports lookup by form (alphabetical headwords) and by function (e.g., request, response, repair-initiation), based on a controlled functional taxonomy aligned with core concepts of Conversation Analysis. The online-resource responds to a documentation gap for German. Lexical units in spoken language, particularly in social interaction, have not yet been brought together in a single resource that serves users with diverse linguistic backgrounds – from linguists to L2-German teachers and students at universities worldwide. The resource will be published online with a restricted number of entries in early 2027, following a user study.

Contact information

Sophia Fiedler
Leibniz-Institut für Deutsche Sprache
sophia.fiedler@ids-mannheim.de

Ivana Filipović Petrović & Slobodan Beliga

MODELLING PHRASEOLOGICAL SEMANTIC STRUCTURES IN DMLEX

Aligning a Croatian Lexonomy-Based Idiom Dictionary with a Thematic Index

Keywords Phraseological units; semantic networks; relational modelling; digital lexicography; DMLex

Phraseological resources contain complex semantic structures that are often difficult to represent in interoperable lexical models. Although digital phraseological dictionaries increasingly use hyperlinks, thematic indices, and semantic groupings to improve navigation, many of these relations remain implicit and are primarily intended for human interpretation. This study explores the expressive capacity of the DMLex model for representing conceptual and semantic relations in a Croatian phraseological resource originally developed in Lexonomy. The resource consists of the *Online Dictionary of Croatian Idioms* and a manually curated thematic index grouping phraseological units into semantic domains and subdomains. The study adopts a scenario-based modelling approach focused on recurring phraseological structures observed in the resource. Three modelling scenarios are analysed: (1) variant and constructional relation networks, (2) polysemy and sense-level thematic classification, and (3) multiple thematic affiliation. The analysis demonstrates how lexical variation, constructional relations, synonymy, and thematic organisation can be represented through explicitly encoded semantic relations within DMLex. The results show that DMLex can accommodate much of the semantic complexity of phraseological resources while also revealing challenges related to overlapping classifications and conceptual hierarchies. The proposed approach contributes to the interoperable modelling of phraseological data for lexicographic and NLP-oriented applications.

Contact information

Ivana Filipović Petrović
Croatian Academy of Sciences and Arts
ivana.filipov@gmail.com

Slobodan Beliga
Faculty of Informatics and Digital Technologies, University of Rijeka
sbeliga@uniri.hr

Hanna Fischer, Alfred Lameli,
Nathalie Mederake & Nico Urbach

LINKING DICTIONARIES THROUGH MEANING

A Sense-Level Evaluation of Taxonomy-Guided LLM Classification in Lexicography

Keywords LLM classification; semantic interoperability; dialect lexicography; onomasiological taxonomy; sense alignment

Despite the growing availability of digital dictionaries, interoperability typically remains limited to headwords, metadata, and formal representation. Senses, by contrast, remain isolated, since dictionary meanings are resource-specific constructs shaped by varying editorial traditions rather than standardized semantic units. This paper evaluates taxonomy-guided LLM classification as a scalable framework for sense-level interoperability across four major German dialect dictionaries, mapping their sense descriptions onto Post's onomasiological taxonomy, which functions as a tertium comparationis for cross-resource sense comparison. Performance is assessed along four dimensions, exact classification correctness, hierarchical proximity, run-to-run reproducibility, and majority-based reliability, with GPT-5 reaching 75–81 % exact correctness and Fleiss' κ between 0.78 and 0.83 across resources. The results show that taxonomy-guided prompting supports robust cross-resource comparability despite orthographic variation, divergent sense granularity, and inconsistent semantic markup. Disagreement patterns prove diagnostically informative rather than merely residual error: we distinguish four sources of uncertainty, stochastic variation, conceptual ambiguity, categorial indeterminacy, and data-related underspecification, each with different implications for lexicographic review workflows. We argue that semantic interoperability in dialect lexicography requires explicit mediating models of meaning, and that taxonomy-guided LLM classification provides a reproducible and interpretable workflow for linking heterogeneous lexical resources at the level of meaning.

Contact information

Hanna Fischer
Forschungszentrum Deutscher Sprachatlas
hanna.fischer@uni-marburg.de

Alfred Lameli
Forschungszentrum Deutscher Sprachatlas
lameli@uni-marburg.de

Nathalie Mederake
Forschungszentrum Deutscher Sprachatlas
mederake@staff.uni-marburg.de

Nico Urbach
Forschungszentrum Deutscher Sprachatlas
nico.urbach@uni-marburg.de

Margarita Freixas, Esther Hernández &
Marija Žarković Eriksson

E-*DRAE* 1884

Towards a Digital and Hypertextual Critical of a Nineteenth-Century Dictionary

Keywords Spanish lexicography; digital humanities; dictionary of the Real Academia Española; nineteenth century

This paper presents the coordinated project "Model for a Digital and Hypertextual Edition of the *DRAE* 1884 (e-*DRAE* 1884)", carried out at the Autonomous University of Barcelona and the Spanish National Research Council (CSIC). Its main goal is to produce an online critical edition of the *Diccionario de la lengua castellana* published by the Real Academia Española in 1884 (hereafter, *DRAE* 1884). The 1884 edition of the *DRAE* was chosen because of its innovative character within the lexicographical tradition of the Real Academia Española and because it was the first edition to take into account the lexical proposals of the newly established American academies. The digital and hypertextual edition offers the full text in an interoperable format, https://edrae1884.uab.cat/edicion/, enriched with structured annotations that provide the information needed for the historiographical, lexicographical, and lexicological interpretation of the dictionary's contents.

References

DRAE 1884 = Real Academia Española (1884). *Diccionario de la lengua castellana*. Imprenta Gregorio Hernando.

e-DRAE 1884 = e-DRAE 1884. https://edrae1884.uab.cat/

Contact information

Margarita Freixas
Universitat Autònoma de Barcelona
Margarita.Freixas@uab.cat

Esther Hernández
Agencia Estatal Consejo Superior de Investigaciones Científicas (CSIC)
esther.hernandez@cchs.csic.es

Marija Žarković Eriksson
Universitat Autònoma de Barcelona
marija.zarkovic@autonoma.cat

Christine Ganslmayer & Martina Nied Curcio

HOW GERMAN LANGUAGE TEXTBOOKS ADDRESS THE DIGITAL TRANSFORMATION OF LEXICOGRAPHY

Keywords Lexicographic competence; digital lexicography; dictionary use; language didactics; German as a foreign language; German as a first language; media literacy

Despite the rapid development of digital lexicographic resources, their integration into German language teaching – both as a first language (L1) and as a foreign language (L2) – remains limited. This paper investigates the discrepancy between current lexicographic practice and language didactics through a qualitative and quantitative analysis of German textbooks (GFL = German as a foreign language: 2011–2024; GL1 = German as a first language: 2003–2021). The findings show that lexicographic resources are underrepresented, predominantly conceptualised in traditional (print-based) ways, and rarely embedded in meaningful communicative or strategy-oriented tasks. Furthermore, the development of lexicographic competence and media literacy is largely neglected. At the same time, learners express a strong demand for guidance in using digital tools, while teachers often lack the necessary training. The paper argues for a systematic reconceptualization of lexicographic didactics, integrating digital resources, hybrid tools, and AI-based applications. It proposes concrete didactic approaches to foster effective, reflective, and critical use of lexicographic resources.

Contact information

Christine Ganslmayer
Friedrich-Alexander-Universität (FAU) Erlangen-Nürnberg
Christine.Ganslmayer@fau.de

Martina Nied Curcio
Università degli Studi Roma Tre
martina.nied@uniroma3.it

Marta Garcia Casado

NEW SENSES, NEW METHODS

Clustering as a Guide for Detecting Semantic Neologisms

Keywords Lexicography; computational semantics; semantic neology; NLP; HDBSCAN

Identifying semantic neology remains a significant challenge for lexicography due to the gradual nature of meaning shifts. This study proposes a methodological framework for the automatic detection of verbal semantic neology in Catalan, leveraging recent advances in Natural Language Processing (NLP). Utilizing a general-language corpus spanning 2000-2021, this research employs a Large Language Model to generate contextualized embeddings for 90 target verbs, capturing nuanced semantic variation. The HDBSCAN clustering algorithm is then applied to group these representations by usage similarity. To validate the results, a lexicographic baseline is established using the *Diccionari essencial de la llengua catalana*; clusters that do not align with registered senses are flagged as candidate neologisms for qualitative analysis. This qualitative analysis focuses on shifts in argument structure and semantic implicatures. The findings emphasize the value of NLP tools for low resource languages like Catalan that have growing but comparatively few digital resources. Ultimately, this study advocates for a hybrid lexicographic model where machine learning serves as an exploratory tool to augment, rather than replace, human expertise. This approach ensures that modern dictionaries can responsibly integrate technological innovation while maintaining academic and ethical standards.

Contact information

Marta Garcia Casado
Universitat Pompeu Fabra
marta.garciac@upf.edu

Pol Garriga Martínez

THE VOCABULARY OF MENTAL HEALTH

Exploitation of Specialized Diachronic Corpora with Artificial Intelligence for Qualitative Semantic Analysis

Keywords Mental health; mental illness; artificial intelligence; AI; hysteria; corpus linguistics; terminology

This article explores the diachronic semantic evolution of mental health vocabulary, focusing specifically on the concept of hysteria across Spanish and German specialised corpora. Drawing on Diachronic Prototype Semantics, the paper examines how semantic changes reflect shifts in radial, hierarchical networks where meanings are updated according to communicative needs. The key methodological aspect is the use of the Gemini large language model to assist in the qualitative analysis of the MentEs and DeMens corpora contexts. This AI-supported approach enables the large-scale categorisation of contexts into six thematic areas: aetiology, clinical manifestations, differential diagnosis, treatment, conceptual debate, and non-clinical uses. The findings indicate that historical scientific discourse was overwhelmingly preoccupied with conceptual delimitation and identification of hysteria (representing 86–94% of occurrences) rather than therapeutic measures. The author concludes that AI-assisted processing facilitates a comprehensive bridge between quantitative data and qualitative interpretation, allowing researchers to document prototypical modulation, for instance, how specific meanings gain or lose prominence depending on the historical and social context.

Contact information

Pol Garriga Martínez
Universitat Pompeu Fabra
pol.garriga@upf.edu

Zoe Gavriilidou

SPANISH LOANWORDS IN MODERN GREEK

Linguistic Analysis and Lexicographic Treatment in the Dictionary of Standard Modern Greek

Keywords Language contact; lexical borrowing; Spanish loanwords; Modern Greek; Greek lexicography

This study aims to provide a systematic linguistic and lexicographic analysis of Spanish-derived loanwords in Modern Greek, addressing the limited attention given to Spanish as a donor language. Methodologically, it draws on a corpus of 130 items, combining 105 entries from the Dictionary of Standard Modern Greek (DSMG) with 25 additional items collected from contemporary digital discourse. The analysis adopts a typological framework distinguishing loanwords, loanblends, and loanshifts, alongside grammatical, morphophonological, and semantic-domain analyses.

The results indicate a clear predominance of loanwords, with loan translations and semantic extensions occurring less frequently. Nouns constitute the dominant grammatical category, while morphophonological integration is largely systematic, aligning with Greek inflectional patterns. Semantically, borrowings cluster in culturally salient domains such as food, music, and material culture, whereas core vocabulary remains largely unaffected. The findings also highlight the predominantly indirect transmission of Spanish borrowings through intermediary languages and the accelerating role of global media in recent diffusion.

The study contributes by documenting Spanish as a secondary but meaningful donor language in Modern Greek and by identifying gaps in the DSMG, advocating for more systematic etymological attribution and a corpus-informed, dynamic approach to lexicographic practice.

Contact information

Zoe Gavriilidou
Democritus University of Thrace
zoegab@otenet.gr

Zoe Gavriilidou & Evanthia Konstantinidou

PEDAGOGICAL LEXICOGRAPHY IN PRACTICE

Raising Dictionary Use Strategies through Card-Based Instruction

Keywords Pedagogical lexicography; dictionary use strategies; vocabulary development; learner autonomy; differentiated instruction; dictionary culture

This paper presents the pedagogical rationale, theoretical foundations, and instructional design of the book *100 Practice Cards for Dictionary Use and Vocabulary Development* (Gavriilidou & Konstantinidou, 2025), a classroom-oriented educational resource developed to promote strategic dictionary use among upper-primary and lower-secondary learners. Grounded in contemporary pedagogical lexicography, strategy-based instruction, and vocabulary acquisition research (Hartmann, 2001; Nation, 2013), the material aims to transform dictionary consultation from a mechanical reference activity into a meaningful learning process that supports vocabulary development, metacognitive awareness, learner autonomy, and language proficiency.

The study highlights the broader educational value of the material, especially its contribution to inclusive education, strategic dictionary use, soft skill cultivation and autonomous language learning.

References

Gavriilidou, Z. & Konstantinidou, E. (2025). *100 καρτέλες εξάσκησης στη χρήση του λεξικού και ανάπτυξης λεξιλογίου*. Δίσιγμα.

Hartmann, R. R. K. (2001). *Teaching and Researching Lexicography*. Pearson Education.

Nation, I. S. P. (2013). *Learning Vocabulary in Another Language* (2nd ed.). Cambridge University Press.

Contact information

Zoe Gavriilidou
Democritus University of Thrace
zoegab@otenet.gr

Evanthia Konstantinidou
Democritus University of Thrace
evi1990@hotmail.gr

Alexander Geyken & Gregor Middell

TREND WORD DETECTION FOR DWDS, A LEXICAL INFORMATION SYSTEM FOR CONTEMPORARY GERMAN

Keywords Trend word detection; keyness metrics; legacy dictionaries; entry revision; lexicographic process

This paper addresses the challenge of prioritizing dictionary revisions in corpus-based lexicography, specifically for the Digital Dictionary of German (DWDS), which relies heavily on legacy content. Given the massive data volume in modern corpora, manual selection of relevant words for updating is no longer feasible. We investigate two methods for identifying "trend words"—terms showing an overproportionally high present-day frequency relative to past intervals. The first method applies established linear regression to large monitor corpora, highlighting limitations due to reliance on the chosen time period. To mitigate this, the second method employs a dispersion-oriented keyness metric on a specialized "discourse corpus" of current online text. This metric scores lemmas based on their spread across unique sources, effectively prioritizing words relevant to contemporary discourse. Evaluation of the ranked candidate list confirms that the high-score group contains a significantly higher density of relevant trend words compared to lower-ranked groups. This system allows lexicographers to systematically prioritize legacy entries for revision, such as those that are outdated or missing a current sense, thereby optimizing the lexicographical workflow and ensuring the dictionary reflects current language use.

Contact information

Alexander Geyken
Berlin-Brandenburg Academy of Sciences and the Humanities
geyken@bbaw.de

Gregor Middell
Berlin-Brandenburg Academy of Sciences and the Humanities
gregor.middell@bbaw.de

Rufus H. Gouws & Theo J.D. Bothma

OPTIMISING INTERFACE DESIGN IN ONLINE LEXICOGRAPHIC PRODUCTS

Keywords Individualisation; interface design; screenshot structure; transformative lexicography

This paper focuses on certain aspects of interface design in online dictionaries. A discussion of a number of examples of interface design in existing online dictionaries is presented. It is emphasised that successful interface design in transformative lexicographic products should be characterised by a human-centred approach that enables intuitive access to the required data. Dictionary articles should offer users an overview of the available contents. The user should then be able to zoom in on specific data types whilst others could be filtered out so that information on demand can be retrieved. An interdisciplinary approach is needed to optimally utilise the possibilities offered by the online environment to design a dynamic article structure with an interface that gives the users rapid access to specific data. Lexicographic theory should offer an exposition of innovative procedures already prevailing in existing dictionaries. Lexicographic theory should also make proposals that can be applied in the practice to enhance the interface design in order to assist users in experiencing successful dictionary consultation endeavours.

Contact information

Rufus H. Gouws
Stellenbosch University
rhg@sun.ac.za

Theo J.D. Bothma
University of Pretoria
theo.bothma@up.ac.za

Barbara Heinisch

DICTIONARIES, GLOSSARIES AND TERMINOLOGY DATABASES IN THE AGE OF ARTIFICIAL INTELLIGENCE

Rethinking Boundaries and Functions

Keywords Lexicography; terminology work; knowledge graph; large language models; retrieval-augmented generation; terminology-augmented generation

The increasing integration of large language models (LLMs) into (multilingual) specialised communication is challenging traditional distinctions between dictionaries, glossaries and terminology databases. While these resources have traditionally differed in structure, purpose and theoretical foundations, LLMs increasingly treat them as machine-accessible sources of lexical knowledge. This paper examines how current approaches, ranging from prompt-based glossary integration and retrieval-augmented generation (RAG) to terminology-augmented generation (TAG) and knowledge graphs, are transforming the use of lexical resources. It argues that reducing rich lexicographic and terminological resources to simple 'equivalence lists' risks losing contextual and conceptual information. At the same time, API-based access to dictionaries and terminology databases enables LLMs to exploit the richness of these resources without prior 'poorification'. The paper proposes shifting the focus from resource typologies to resource functions. It explores how lexical knowledge can be structured, accessed and governed for use in LLMs, and concludes that lexicographers and terminologists are evolving into knowledge engineers who mediate between language, knowledge and language technologies.

Contact information

Barbara Heinisch
Eurac Research
barbara.heinisch@univie.ac.at

Ondřej Herman

SENSE CLASSIFICATION FOR WORD SKETCHES

Keywords Word sense induction; Word Sketch; polysemy; corpus lexicography; Sketch Engine

The Word Sketch is a core tool for modern corpus-based lexicography, providing a concise summary of a word's grammatical and collocational behavior. A longstanding limitation of standard Word Sketches is the lack of polysemy awareness: when a lexicographer examines a highly polysemous word such as *bank*, the resulting collocate lists mix financial terms like *account* and *deposit* with geographical terms like *river* and *sand*. In this paper, we present and evaluate a fully automated method for enriching Word Sketches with word sense information. Using the Adaptive Skip-gram model, we perform unsupervised word sense induction and project the induced senses onto the type-level collocate lists of a Word Sketch, so that each collocate is associated with the sense it most strongly represents. On a manually annotated test set of 23 polysemous English words, our method achieves a Precision of 0.64 and Recall of 0.77 (F1 $\approx$ 0.70). The method requires no manual training data and is already deployed in Sketch Engine for eight languages. We describe the user interface, which allows lexicographers to filter Word Sketches by sense, and discuss implications for dictionary-writing workflows.

Contact information

Ondřej Herman
Lexical Computing
ondrej.herman@sketchengine.co.uk

Polina Iaroshenko & Natalia Loukachevitch

MORPHOLOGICAL FORMS IN EMOTION LEXICONS

Evidence from LLMs

Keywords Emotions; large language models; lexicon; natural language processing

Emotion expression at the lexical level has been extensively studied, whilst considerably less research addresses emotion expression through morphological devices. This study demonstrates that different morphological forms of the same lexical unit may convey different emotions, as revealed through experiments with large language models (LLMs). Material was drawn from RusEmoLex, a Russian emotion lexicon comprising 1,024 lexical units across five emotion categories (joy, sadness, anger, fear, surprise). For 473 verbs, 12 morphological forms were generated (5,676 in total); for 336 nouns, 6 case forms were generated (2,016 in total). Eight LLMs (7B–671B parameters) classified each form. Performance was evaluated using weighted F1-score. Results reveal substantial asymmetry between word classes. Noun case forms showed minimal morphological influence (mean difference: 3.6%), whilst verb forms demonstrated considerably greater variation (11.5%). Past tense feminine forms consistently achieved the best performance across all models; imperative forms systematically underperformed. A dominant error pattern emerged: sadness-labelled verbs in imperative forms were frequently misclassified as anger. Semantic analysis suggests this pattern may arise from interaction between grammatical form and lexical semantics, with imperative forms potentially amplifying violence-associated meanings in agentive verbs. Findings indicate that incorporating morphological variants may enhance emotion lexicon effectiveness for morphologically rich languages.

Contact information

Polina Iaroshenko
Lomonosov Moscow State University
polina.iaroshenko@yandex.ru

Natalia Loukachevitch
Research Computing Center of Moscow State University
louk_nat@mail.ru

Ellert Thor Johannsson, Jóhannes B. Sigtryggsson &
Thordis Ulfarsdottir

UNLOCKING A NINETEENTH CENTURY DICTIONARY

A Closer Look at Konráð Gíslason's *Dönsk orðabók* (1851)

Keywords Lexicography; language purism; digitisation; historical lexicography

This article discusses an important Icelandic lexicographic work published in the nineteenth century. *Dönsk orðabók með íslenzkum þýðingum* was compiled and edited by Konráð Gíslason (1808-1891), an Icelandic academic and language purist. The printed work is described in some detail, explaining the editorial principles and how purist ideas shaped the form of many of the definitions where loanwords were avoided and paraphrasing was preferred when the target language did not have a native equivalent. The discussion then moves on to recent efforts to produce a digitised version of the dictionary. We account for the workflow implemented in order to turn this complicated printed artifact into a functioning electronic edition where the content has been marked-up and made searchable. The described process illustrates some of the issues involved and addresses challenges encountered when producing the new digital edition of the dictionary.

Contact information

Ellert Thor Johannsson
The Árni Magnússon Institute for Icelandic Studies
etj@hi.is

Jóhannes B. Sigtryggsson
The Árni Magnússon Institute for Icelandic Studies
johans@hi.is

Thordis Ulfarsdottir
The Árni Magnússon Institute for Icelandic Studies
disa@hi.is

Boris Kern

WORD-FORMATION CHAINS IN SLOVENIAN

Evaluating the Effectiveness of AI

Keywords Multistage word formation; morphotactics; formants; AI

This study investigates word-formation chains, i.e. sequences of derivationally related words in which each derivative serves as the base for the next. Due to their rich morphemic structure, such chains are particularly complex in Slavic languages. Their analysis is relevant for understanding word-formation and semantic potential, morphotactics, and lexicographic criteria for vocabulary inclusion. The research builds on the project *Formant Combinatorics in Slovenian* (2022–2025), which applied automatic extraction methods to data from the *Word-Family Dictionary of the Slovene Language* (BSSJ). Abstract suffix-chain patterns derived from this resource were extended to larger lexical datasets. Preliminary findings showed limited success of automatic extraction methods (24% for the most frequent suffix chain), whereas AI achieved higher accuracy (59%). However, performance declined markedly with increasing chain complexity. The present study aims to evaluate the effectiveness of ChatGPT 5.5 in identifying word-formation chains with nominal, verbal and adjectival bases. It focuses on the impact of chain complexity, the role of part-of-speech within chains, and problematic suffixes and zero-formant derivation patterns. Particular challenges include derivatives without overt suffixes and cases where zero-formant nouns are misclassified as unmotivated words. The analysis examines suffix chains of varying complexity and compare AI performance with previous automatic extraction approaches, with the goal of improving both extraction methods and AI prompting strategies.

Contact information

Boris Kern
Research Centre of the Slovenian Academy of Sciences and Arts; University of Nova Gorica
boris.kern@zrc-sazu.si

Václava Kettnerová, Veronika Kolářová & Jiří Mírovský

WHERE THE BOUNDARY BETWEEN DERIVATIONAL AND INFLECTIONAL MORPHOLOGY MEETS LEXICOGRAPHY

Building the DeriVallex Lexicon

Keywords Valency; valency lexicons; derivation; inflection; word-formation categories; transposition; noun and adjectival derivatives; verb forms; noun and adjectival lemmas

The lexicographic treatment of word-formation categories that lie at the boundary between inflection and derivation remains an open question. In this paper, we introduce the process of compiling *DeriVallex*, a valency lexicon containing automatically generated valency frames, which provides information on the valency behavior of Czech noun and adjectival deverbal derivatives being halfway between forms and lexemes. These word-formation categories inherit their valency frames and other valency-related phenomena (as, e.g., reflexivity, reciprocity, and other semantic alternations) from their base verbs. We first outline the automatic generation of valency frames of these noun and adjectival derivatives, which resulted in 17,145 valency frames assigned to lexical units, i.e., individual senses of these derivatives, represented by 11,199 noun and adjectival lemmas. The main focus of this paper in on the process of structuring of the obtained data into a valency lexicon, providing rich information on valency and derivational properties of the selected nouns and adjectives, enabling human users to search and filter the data. The resulting data may be integrated into both valency lexicons and explanatory dictionaries, thereby increasing their descriptive comprehensiveness.

Contact information

Václava Kettnerová
Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics
kettnerova@ufal.mff.cuni.cz

Veronika Kolářová
Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics
kolarova@ufal.mff.cuni.cz

Jiří Mírovský
Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics
mirovsky@ufal.mff.cuni.cz

Agnes Kim

DIGITAL METALEXICOGRAPHY AS A TOOL FOR HISTORICAL (CONTACT) LINGUISTICS

Keywords Viennese; Slavic languages; Czech; metalexicography; language contact; loan words

This contribution presents a data collection on 'Slavic in Viennese', i.e. (glottonyms for) Slavic languages, (ethnonyms for) their speakers as well as single linguistic phenomena in German varieties typical for Vienna, that comprises 6,406 articles on 268 lemmas from 30 dialect dictionaries published between 1800 and 2020. In the contribution's first part, corpus compilation and the underlying data structure are described and results of existing studies on this data collection are discussed. So far, these have allowed for the identification of reception networks between the dictionaries and revealed the reconstruction of stereotypes regarding speakers of South Slavic languages by them. Inspired by these results, the second part of the contribution presents an approach towards the use of the data collection (and similar ones) for the description of integration, variation and change of (West Slavic) loanwords in 'Viennese'. By means of exploratory cluster analyses, three main clusters of lemma-meaning pairs are identified: one general that includes forms and meanings used throughout the complete observation period, one of historical forms and meanings that were common in the 19th century, and one of innovations from the second half of the 20th century. Building on this and the exemplary comparison of forms and meanings of three Czech loanwords (*Kaluppe*, *Kolatsche*, *pomali*), the contribution suggests that most similar loanwords must have been borrowed before the 19th century, thereby demonstrating the usability of the approach for historical (contact) linguistics.

Contact information

Agnes Kim
University of Ostrava
agnes.kim@osu.cz

Veronika Kolářová, Václava Kettnerová,
Michal Olbrich & Jiří Mírovský

NOMVALLEX

Valency of Czech Nouns and Adjectives in Networks of Derivationally Related Words

Keywords Valency lexicon; nouns; adjectives; changes in valency; derivational relation; word-formation network; graph representation

NomVallex is a valency lexicon of Czech nouns and adjectives, aimed primarily at describing valency of individual noun and adjectival derivational categories in response to their insufficient theoretical description and coverage in lexical resources. Annotation has proceeded from the more extensively studied derivational categories (deverbal nouns and adjectives) through deadjectival nouns to the least researched denominal derivatives and compounds. Since valency is often affected by word-formation relations, NomVallex has implemented features for capturing derivational relations between derivationally related lexical units, and for comparing their valency. Although the annotation in NomVallex often encompasses more than one derivative of a particular base word, only a limited part of the word-formation relations in Czech is covered there. The lexicon therefore newly provides links to DeriNet, where users can view a given word in the broader context of its derivational relations. Moreover, to display derivational relations and changes in valency associated with the word-formation process in whole networks of derivationally related words recorded in NomVallex, version 2.7 has introduced a DeriNet-like graphical representation of derivational relations, which is available in a user-friendly format on the lexicon's website, facilitating further linguistic research into the influence of word formation on the valency of derived words.

Contact information

Veronika Kolářová
Institute of Formal and Applied Linguistics, Charles University, Faculty of Mathematics and Physics
kolarova@ufal.mff.cuni.cz

Václava Kettnerová
Institute of Formal and Applied Linguistics, Charles University, Faculty of Mathematics and Physics
kettnerova@ufal.mff.cuni.cz

Michal Olbrich
Institute of Formal and Applied Linguistics, Charles University, Faculty of Mathematics and Physics
olbrich@ufal.mff.cuni.cz

Jiří Mírovský
Institute of Formal and Applied Linguistics, Charles University, Faculty of Mathematics and Physics
mirovsky@ufal.mff.cuni.cz

Mariia Kopylova

AI-ASSISTED DETECTION AND MITIGATION OF LEXICOGRAPHIC BIAS

Semantic Reduction in Definitions of Five Socio-Legal Terms

Keywords Lexicographic bias; semantic reduction; Sentence-BERT; cosine similarity; NLP; legal lexicography

When non-specialist users consult general dictionaries to understand legally significant terms, they risk encountering definitions that are fundamentally incomplete – without any indication of this gap. This study introduces a reproducible computational methodology to quantify this semantic reduction and explore AI tools for its mitigation.

Five socio-legal terms were examined: *harassment*, *stalking*, *discrimination*, *abuse*, and *neglect*. A general corpus of 35 definitions was compiled from the Cambridge Dictionary, Oxford English Dictionary, and Merriam-Webster; a legal corpus of 15 definitions was drawn from Black's Law Dictionary, Oxford Dictionary of Law, and Merriam-Webster's Dictionary of Law. Definitions were vectorised using Sentence-BERT and compared via cosine similarity.

None of the five terms reached the 0.70 similarity threshold. Mean similarity ranges from 0.491 (abuse) to 0.571 (stalking); 92.4% of all pairwise comparisons fall below this threshold. Statistical significance was confirmed by one-sample t-tests ($p < 0.01$). The Qwen2.5-3B-Instruct model extracted key legal components for each term, which can serve as editorial checklists for lexicographers reviewing dictionary entries.

The findings demonstrate that semantic reduction is systematic, measurable, and varies across terms. AI tools can assist lexicographers in detecting and addressing such gaps, though final decisions remain with the human expert.

Contact information

Mariia Kopylova
Roma Tre University
mashulya.kopylova@list.ru

Katharina Korecky-Kröll, Philipp Stöckle, Daniel Elsner, Wolfgang Koppensteiner & Elisabeth Eder

LLM-ASSISTED LEXICOGRAPHY IN LOW-RESOURCE CONTEXTS

From Prompt to Dictionary Article

Keywords LLM; low-resource languages; dictionary articles; German; prompt engineering

This paper investigates the potential of Large Language Models (LLMs) to support lexicographic work in low-resource contexts, focusing on Austrian Bavarian dialects documented in the *Wörterbuch der bairischen Mundarten in Österreich* (*WBÖ*). Using a structured prompt-engineering workflow and a three-stage data pipeline, 100 dictionary articles were generated with LLaMA 4 (Scout) and systematically compared to human-authored "gold" articles. Evaluation combined automated metrics, LLM-based judgement, and assessment by three human lexicographers. Results show that LLMs reliably reproduce formal lexicographic conventions, producing well-structured and coherent dictionary entries. Nevertheless, limitations emerge in semantic completeness and data fidelity: generated articles sometimes omit senses or misrepresent semantic structure and often inadequately integrate empirical evidence, especially regional information and attestations. While automated metrics indicate a relatively high similarity of original and generated senses based on their embeddings, both human evaluators and LLM-as-a-judge approaches reveal persistent weaknesses. However, the strong alignment between human judgements and the sense similarity metric suggests that embedding-based evaluation is a promising proxy for human assessment, whereas holistic LLM-as-a-judge scores appear comparatively strict and less transparent. The study argues for a hybrid evaluation framework and positions LLMs as assistive tools within human-in-the-loop lexicography.

Contact information

Katharina Korecky-Kröll
Austrian Academy of Sciences
katharina.korecky-kroell@oeaw.ac.at

Philipp Stöckle
Austrian Academy of Sciences
philipp.stoeckle@oeaw.ac.at

Daniel Elsner
Austrian Academy of Sciences
daniel.elsner@oeaw.ac.at

Wolfgang Koppensteiner
Austrian Academy of Sciences
wolfgang.koppensteiner@oeaw.ac.at

Elisabeth Eder
Austrian Academy of Sciences
elisabeth.eder@oeaw.ac.at

Iztok Kosem, Polona Gantar, Simon Krek & Tjaša Arčon

DIVIDE AND DEFINE WITH LLMS

Generating Definitions for the Digital Dictionary Database of Slovene

Keywords LLM; definitions; dictionary database; evaluation; Slovene

This paper presents an experiment on generating definitions for Slovene headwords using four Large Language Models (LLMs): two commercial (Gemini and GPT) and two open-source (Gemma and the Slovenian model GaMS). Each model was provided with contextual data, including semantic indicators, collocations, and examples. We tested a zero-shot approach and two few-shot approaches using sample definitions from the Digital Dictionary Database of Slovene (DDDS) and the New Dictionary of Standard Slovene. The results show that while all models produce high-quality definitions, Gemini performs best across all word classes, particularly when utilizing DDDS sample definitions. The second part of the study evaluated LLMs as judges of definition quality, testing Gemini, GPT, and Claude-Sonnet. Gemini again achieved the highest agreement with human lexicographers. Notably, even when LLMs selected a different definition than the experts, their choices were usually viable. These results demonstrate the considerable potential of LLMs in lexicography, and we plan to deploy this pipeline to generate definitions for a much larger set of Slovenian headwords.

Contact information

Iztok Kosem
Faculty of Arts, University of Ljubljana
iztok.kosem@cjvt.si

Polona Gantar
University of Ljubljana
apolonija.gantar@guest.arnes.si

Simon Krek
Jožef Stefan Institute
simon.krek@guest.arnes.si

Tjaša Arčon
Faculty of Computer and Information Science, University of Ljubljana
tjasa.arcon@fr.uni-lj.si

František Kovařík, Marek Blahuš, Miloš Jakubíček & Vojtěch Kovář

AGREE TO DISAGREE

Intra-Annotator Agreement in Semi-Automatic Vocabulary Building

Keywords Inter-annotator agreement; intra-annotator agreement; dictionary drafting; post-editing lexicography; corpus annotation; semi-automatic dictionary making; Dictionary Express; Czech

Intra-annotator agreement is the rate at which an annotator makes the same choices in the same situation. This paper examines the annotation process in a semi-automatic dictionary-making project, Czech Dictionary Express. In the vocabulary-building process, its annotators went through 100,000 headwords from a corpus frequency wordlist, marking each as incorrect or (partially) correct. The aim of the Dictionary Express projects is a high dictionary-making speed, so the annotators weren't given the context of the headwords.

In this paper, we examine the shift in annotators' opinions on headwords before and after context was provided. We argue against providing no context of the headwords, because, unlike the (partially) accepted words, which can still be cut from the vocabulary at later stages, the rejected words cannot be put back in, and this is the case of some correct words that have not been recognised by the annotators without context. We show that the context-based annotation process doesn't take significantly longer. We also take a brief look at a study about using LLMs for the same annotation process.

Contact information

František Kovařík
Lexical Computing / Faculty of Arts, Masaryk University, Brno, Czech Republic
frantisek.kovarik@sketchengine.eu

Marek Blahuš
Lexical Computing
marek.blahus@sketchengine.eu

Miloš Jakubíček
Lexical Computing / Faculty of Informatics, Masaryk University, Brno, Czech Republic
milos.jakubicek@sketchengine.eu

Vojtěch Kovář
Lexical Computing / Faculty of Informatics, Masaryk University, Brno, Czech Republic
vojtech.kovar@sketchengine.eu

Dominika Kováříková & Laura A. Janda

EMOTIONS IN CONTEXT

A Semantic Field Viewed through Grammar and Collocations

Keywords Semantic field; emotion nouns; grammatical profiles; collocations

We examine Czech emotion nouns as a structured semantic field rather than as a set of isolated dictionary entries. The analysis is based on 65 nouns denoting emotions, selected from the Czech reference corpora SYN2015 and SYN2020. It combines grammatical profiling, correspondence analysis, and a brief collocational overview in order to show that these lexemes share certain patterns, some of them characteristic of the field as a whole and others pointing to its internal structuring. The results suggest that grammatical behaviour does not simply reflect semantic similarity in the usual psychological sense, but reveals different ways in which emotions are represented in language, for example as experienced states, emotional contents, agent-like forces, or states one is driven into. Collocational evidence points to broader tendencies across the field, including loss of agency, emotional intensity, bodily manifestation, regulation, and combination with other emotions. We argue that this kind of field-based description can complement traditional lexicographic practice and may also be attractive to dictionary users interested in contrasts between related words.

Contact information

Dominika Kováříková
Charles University
dominika.kovarikova@ff.cuni.cz

Laura A. Janda
The Arctic University of Norway
laura.janda@uit.no

Barbara Lewandowska-Tomaszczyk

THE INTEGRATION OF BRITISH AND AMERICAN SLANG IN CONTEMPORARY POLISH AND THE ROLE OF LLM-DRIVEN DYNAMIC LEXICOGRAPHY

Keywords American English; borrowings; British English; Dynamic Lexicography; LLMs; Polish; semantic shifts; slang; Pilot Study; limitations

This paper investigates the accelerating influx of British and American English slang into the Polish language, driven by digital globalization and social media language use. While traditional Anglicisms often focus on technical or professional domains, modern borrowing is increasingly characterized by the basic realm of current English loanwords i.e., affective slang—informal expressions of identity, emotion, and subcultural belonging. Central to this research is the introduction of LLM-based (Large Language Model) Dynamic Lexicography as an assisting methodological solution to a traditional lexicography, in which an editorial process fails to capture the rapid evolution and eventual obsolescence of slang terms. This paper proposes a framework utilizing LLMs to perform real-time analysis and mapping across vast corpora of Polish social media data. Such a dynamic, LLM-prompted dictionary might be under constant scrutiny both by volunteer editors and online communities with the existing resources and, as proposed here, results of surveys obtained from Polish youth-based studies.

Contact information

Barbara Lewandowska-Tomaszczyk
University of Applied Sciences Konin
barbara.lewandowska-tomaszczyk@konin.edu.pl

David Lindemann

DEVELOPING A DIGITAL VERSION OF GIDOR BILBAO'S LATIN–BASQUE DICTIONARY

Keywords Dictionary Digitization; Linked Data; Ontolex; Wikibase; Latin; Basque

In this article, we describe several steps taken to convert Gidor Bilbao's *Oinarrizko Hiztegia Latina-Euskara* (Basic Latin–Basque Dictionary) from a paper-based resource and move toward presenting it as Linked Data. Starting from a document in MS Word format, we attempted to isolate and tag elements of the dictionary's macro- and microstructures. The goal of the experiments outlined here was to capture the content of those elements based on the Linked Data paradigm, using the Ontolex-Lemon model. Once integrated into Wikibase, the data are ready for final structuring or further enrichment.

Contact information

David Lindemann
EHU University of the Basque Country
david.lindemann@ehu.eus

Ivana Lovrić Jović & Lana Hudeček

SYNONYMS THAT NEVER MET

On the Compilation of the Dictionary of the Dubrovnik Idiom

Keywords Born digital dictionary; Dubrovnik idiom; historical lexicography; synonymy

This article outlines the Dictionary of the Dubrovnik Idiom project at the Institute for the Croatian Language. Given its unconventional conceptual framework, this dictionary adopts a diachronic approach, incorporating headwords from diverse historical periods (spanning from the 16th to the 21st century) and across a broad spectrum of linguistic registers: from the vernacular to high-prestige literature. From the perspective of a fixed point on the diachronic axis (e.g. present-day usage), some synonyms may, depending on their temporal distance, be classified as obsolete or archaic. However, when arranged along both the synchronic and diachronic dimensions, synonym sets can be further differentiated according to functional-stylistic and register-based distribution. The paper presents a typology of the synonymy encountered during the lexicographic processing. Particular attention is paid to cases in which the dictionary records *synonyms that never met*, that is, synonyms for which corpus evidence shows no temporal overlap.

Contact information

Ivana Lovrić Jović
Institute for the Croatian Language
ilovric@ihjj.hr

Lana Hudeček
Institute for the Croatian Language; Vern' University
lhudecek@ihjj.hr

Tinatin Margalitadze, Giorgi Meladze, Ketevan Mchedlishvili, Tamar Laluashvili & Giorgi Okropiridze

EFFECTIVENESS OF AI FOR THE LEXICOGRAPHY OF LOW-RESOURCE LANGUAGES

The Dictionary of Georgian Neologisms

Keywords Dictionary of neologisms; Georgian language; AI; prompts; lexicographic

Until relatively recently, societies expressed serious concerns about AI and the potential dangers it might pose to humanity. However, this attitude has gradually been replaced by a more pragmatic perspective, as scientific research increasingly focuses on identifying effective ways to harness the opportunities offered by this technology. Lexicographers likewise responded to the advent of AI with numerous studies published, exploring various aspects of AI use in dictionary-making (Trap-Jensen, 2023; Bothma & Gouws, 2025; de Schryver, 2025, *etc.*). Research studies have also been conducted for the Georgian language (Mchedlishvili, 2025; Tchighladze, 2025; Lomidze, 2025).

The purpose of this paper is to present the results of a study that has investigated the effectiveness of AI in the compilation of the *Dictionary of Georgian Neologisms*, an ongoing project at the Centre for Lexicography and Language Technologies of Ilia State University. The experiments related to this research were conducted with ChatGPT 5.2, Gemini 3.0 Thinking, and Grok 4.1 Thinking in combination, in order to mitigate model-specific bias in Georgian.

References

Bothma, Th. & Gouws, R. (2025). Generative Artificial Intelligence (GenAI) as Information Tool for Lexicographic Information Needs. *Lexikos*, 35(2), 276–295. https://doi.org/10.5788/35-2-2092

de Schryver, G.-M. (2025). Customised GPTs for Lexicography. *Lexikos*, 35(2), 397–422. https://doi.org/10.5788/35-2-2104

Lomidze, K. (2025). AI in Lexicography for Low-Resource Languages: The Case of the Georgian Language. In T. Margalitadze (Ed.), *Book of abstracts of the II International Conference Lexicography in the XXI Century* (pp. 97–98). Ilia State University. Available at: https://lexicography21.iliauni.edu.ge/en/book-of-abstracts-2025-2 [accessed 27/03/2026].

Mchedlishvili, K. (2025). Harnessing AI for Creating Definitions of Georgian Military Terminological Neologisms in the context of the Russia-Ukraine War. In T. Margalitadze (Ed.), *Book of abstracts of the II International Conference Lexicography in the XXI Century* (pp. 112–113). Ilia State University. Available at: https://lexicography21.iliauni.edu.ge/en/book-of-abstracts-2025-2/ [accessed 27/03/2026].

Tchighladze, K. (2025). Integrating AI into the Term Creation Process: Case Study of English and Georgian Terms of Emerging Technologies. In T. Margalitadze (Ed.), *Book of abstracts of the II International Conference Lexicography in the XXI Century* (pp. 149–151). Ilia State University. Available at: https://lexicography21.iliauni.edu.ge/en/book-of-abstracts-2025-2/ [accessed 27/03/2026].

Trap-Jensen, L. (2024). The Best of Two Worlds: Exploring the Synergy between Human Expertise and AI in Lexicography. In T. Margalitadze (Ed.), *Proceedings of the I International Conference Lexicography in the XXI Century* (pp. 26–39). Ilia State University. Available at: https://lexicography21. iliauni.edu.ge/wp-content/uploads/2024/06/03_Lars-Trap-Jensen.pdf [accessed 21/02/2026].

Contact information

Tinatin Margalitadze
Ilia State University
tinatin.margalitadze@iliauni.edu.ge

Giorgi Meladze
Ilia State University
giorgi.meladze.4@iliauni.edu.ge

Ketevan Mchedlishvili
Ilia State University
ketevani.mchedlishvili.3@iliauni.edu.ge

Tamar Laluashvili
Ilia State University
tamar.laluashvili.1@iliauni.edu.ge

Giorgi Okropiridze
Ilia State University
giorgi.okropiridze.4@iliauni.edu.ge

Aleksandra Marković, Cvetana Krstev,
Milica Dinić Marinković & Ranka Stanković

TOWARDS A REPOSITORY OF SIMILES IN CONTEMPORARY SERBIAN

Keywords Similes; repository; Serbian language; corpus linguistics; semantic annotation; lexical resources; UNITEX; LLM; NLP

This paper presents the methodological framework for building the *Similes Repository for Contemporary Serbian* (SSR), a corpus-based linguistic resource designed for the collection, annotation, and analysis of similes in contemporary Serbian. The study presents a multi-stage extraction and validation pipeline based on the genre-balanced subcorpus of the *Reference Corpus of Contemporary Serbian*, combining CQL queries, local grammars implemented in UNITEX, and LLM-assisted semantic annotation. The initial extraction phase produced 574,195 candidate sentences across six major simile patterns, confirming both the productivity and structural diversity of simile constructions in Serbian. Subsequent rule-based filtering and LLM-assisted validation substantially reduced noise while preserving a large pool of linguistically relevant figurative comparisons. Manual evaluation performed on a stratified sample additionally revealed that simile interpretation frequently involves distributed semantic structure, and complex interactions between grammatical realization and figurative meaning, and highlighted the importance of linguistic post-processing, as well as several challenges related to semantic annotation in morphologically rich languages such as Serbian. By combining rule-based methods, LLMs, and detailed linguistic analysis, the paper contributes to both the development of computational resources for Serbian and broader research on figurative language annotation and representation.

Contact information

Aleksandra Marković
Institute for Serbian Language, Serbian Academy of Sciences and Arts
aleksandra.markovic@isj.sanu.ac.rs

Cvetana Krstev
JeRTeh Society
cvetanajk@gmail.com

Milica Dinić Marinković
University of Belgrade – Faculty of Philology
milica.dinic.marinkovic@fil.bg.ac.rs

Ranka Stanković
University of Belgrade – Faculty of Geology and Mining
ranka.stankovic@rgf.bg.ac.rs

Peter Meyer

AN AI-DRIVEN NATURAL LANGUAGE INTERFACE FOR GRAPH DATABASE QUERIES IN A LEXICOGRAPHICAL RESOURCE

Keywords Loanword lexicography; semantic parsing; natural-language interface; graph database; text-to-JSON; large language models; query generation

Complex digital lexicographical resources often feature rich data structures that are difficult for non-specialist users to query. This paper presents a pilot study on an LLM-driven natural-language interface for the *Lehnwortportal Deutsch*, a graph-based database of German loanwords in other languages. Rather than translating user questions directly into the database query language, the proposed architecture maps natural-language requests to the JSON configuration format already used by the portal's visual query builder. This intermediate representation makes it possible for users to inspect and modify query builder configurations generated from their questions. We discuss the advantages of this approach and present a pilot study that evaluated it on a set of 310 German questions that, due to lack of real user questions, had to be AI-generated using several different strategies to cover a large range of usage scenarios as well as available querying options. Using one-shot prompting, a state-of-the-art commercial-grade LLM produced valid JSON in almost all cases; 260 outputs were rated perfect and 23 acceptable by human assessment under a low reasoning budget. A high-budget rerun improved most initially bad cases, though with impractical latency for real-time use. The paper also discusses methodological challenges in generating and evaluating synthetic query sets, handling unsupported user requests, and extending the approach through larger datasets, comparing multiple language models, varying output parameters, and using fine-tuning as an alternative strategy.

Contact information

Peter Meyer
Leibniz-Institut für Deutsche Sprache (IDS)
meyer@ids-mannheim.de

Ana Mihaljević & Josip Mihaljević

ARTIFICIAL INTELLIGENCE AS SUPPORT FOR THE DIGITIZATION AND COMPILATION OF THE *DICTIONARY OF THE CROATIAN REDACTION OF CHURCH SLAVONIC*

Keywords Croatian Church Slavonic; large language models (LLMs); digitisation; dictionary compilation; low-resource languages

The *Dictionary of the Croatian Redaction of Church Slavonic* (RCJHR) is the first lexicographic description of the Croatian Church Slavonic language. Within the project *Development of the Digital Infrastructure Model of the Old Church Slavonic Institute* (DigiSTIN), the printed RCJHR is being digitised and adapted for online publication, while the RCJHR project focuses on compiling new entries and exploring a fully digital lexicographic workflow. This paper examines the potential of large language models (LLMs) to support the digitisation and compilation of a dictionary documenting a low-resource historical language with an exceptionally complex microstructure involving five languages and four scripts. Six leading models (GPT-5.4 Thinking, DeepSeek-V3.2, Gemini 3 Pro, Claude Opus 4.6, Qwen3.6-Plus, and Copilot GPT-5) were evaluated across four tasks: OCR of previously published fascicles, HTR of handwritten index cards, defining entry structure, and compiling new dictionary entries. The results demonstrate that LLMs significantly accelerate digitisation and assist in identifying foreign-language equivalents, suggesting sense divisions, and proposing cross-references. However, they remain unreliable in resolving subtle semantic distinctions and recognising the Glagolitic script, confirming that human philological expertise remains indispensable.

Contact information

Ana Mihaljević
Old Church Slavonic Institute, Zagreb
amihaljevi@gmail.com

Josip Mihaljević
Old Church Slavonic Institute, Zagreb
jmihaljevic@stin.hr

Tibor M. Pintér, Manana Rusieshvili-Cartledge, Dóra Pődör, Katalin P. Márkus & Marine Makhatadze

SUBLEXICAL SYSTEMS OF COLOUR IN BIBLICAL GEORGIAN AND HUNGARIAN

Findings of the Pentateuch

Keywords Historical lexicography; cognitive linguistics; contrastive colour semantics; parallel corpora; bible translation; digital lexicography

As a culturally embedded linguistic category, colour varies across translations of the same source text. This paper examines how colour is encoded diachronically in biblical Hungarian and Georgian, addressing two questions: how colour terms are realised across parallel translations, and what cross-linguistic patterns of lexical and syntactic variation emerge. The analysis draws on a parallel corpus of Old and Modern Georgian and Hungarian Bible translations. Colour terms are defined as lexical units denoting chromatic properties, including both basic colour adjectives and material-based or metonymic expressions. The findings reveal a systematic divergence between the two cultural-linguistic traditions. Georgian shows a diachronic shift from materially grounded, periphrastic expressions towards more abstract and generalised colour terminology, often accompanied by increased morphological transparency. Hungarian, by contrast, demonstrates greater lexical continuity, with variation primarily realised through syntactic explicitation and the redistribution of semantic content. In both cultural-linguistic traditions, colour is encoded via material reference, emerging in descriptions of textiles and ritual objects. The findings demonstrate that colour encoding reflects broader processes of semantic restructuring. Hungarian serves as a comparative baseline, highlighting a trajectory in which syntactic rather than lexical change predominates.

Contact information

Tibor M. Pintér
Department of Hungarian Linguistics at Károli Gáspár University of the Reformed Church in Hungary
m.pinter.tibor@kre.hu

Manana Rusieshvili-Cartledge
English Philology at Ivane Javakhishvili Tbilisi State University
manana.ruseishvili@tsu.ge

Dóra Pődör
Department of English at Károli Gáspár University of the Reformed Church in Hungary
podor.dora@kre.hu

Katalin P. Márkus
Department of English at Károli Gáspár University of the Reformed Church in Hungary
p.markus.kata@kre.hu

Marine Makhatadze
English Philology at Ivane Javakhishvili Tbilisi State University
marine_makhatadze@yahoo.com

Bruno Nahod & Marijana Tuta Dujmović

FROM TEXT TO TERMS

TermScout, an AI Module for Early Terminology Processing

Keywords Terminology management; candidate-term identification; large language models; bilingual terminology; AI-assisted terminology processing

This paper presents TermScout, a GPT-based tool developed to support the initial stage of terminology management by identifying candidate terms in specialised texts. The study evaluates TermScout using a corpus focused on cochlear implantation, assessing both English term extraction and the generation of Croatian equivalents. TermScout identified 100 English candidate terms from a specialised textbook, while Croatian equivalents were generated either directly by TermScout or through a subsequent process utilising TermAI. An independent term validation tool was used to assess all results. The tool demonstrated strong performance in English term extraction, achieving a validity rate of 92%. In contrast, generating Croatian equivalents proved more challenging, with validity rates of 22% for TermScout and 29% for TermAI. These findings indicate that concept identification and bilingual term generation are distinct tasks that require distinct linguistic and terminological expertise. The study underscores the potential of AI-assisted workflows for terminology management, while emphasising the continued need for expert validation.

Contact information

Bruno Nahod
Institute of Croatian Language and Linguistics
bnahod@ihjj.hr

Marijana Tuta Dujmović
Polyclinic for Hearing and Speech Rehabilitation SUVAG
mtuta@suvag.hr

Gunta Nešpore-Bērzkalne, Mikus Grasmanis & Madara Stāde

POSSIBILITIES OF REPRESENTATION OF MULTI-WORD EXPRESSIONS IN THE LATVIAN ELECTRONIC DICTIONARY *TĒZAURS*

Keywords MWE entries; electronic dictionary Tēzaurs; intra-dictionary links

This paper aims to examine the existing representation of multi-word expression (MWE) entries in the electronic Latvian dictionary *Tēzaurs* and characterise their interaction with other, single-word entries, as well as highlight future improvement for better integration of MWEs in the dictionary data. The study arises from the need to examine MWE entries in more detail, as they are a substantial part of a language, but are often given less attention than single-word entries in a dictionary. The paper gives an overview of MWE representation within the *Tēzaurs* structure, e.g., links with component-word entries, corpus examples, sources, as well as semantic links with single-word senses or other MWEs. Linking to other resources, such as the Princeton English WordNet is also mentioned. This overview shows that MWEs in *Tēzaurs* function well as independent units that can freely interact with other entries and be embedded in other entries. MWEs work well in the resource's system of links between entries. The next step is expanding the representation of MWEs in this regard.

Contact information

Gunta Nešpore-Bērzkalne
Institute of Matematics and Computer Science, University of Latvia
gunta.nespore@gmail.com

Mikus Grasmanis
Institute of Matematics and Computer Science, University of Latvia
mikus.grasmanis@lumii.lv

Madara Stāde
Institute of Matematics and Computer Science, University of Latvia
stade.madara@gmail.com

Sanni Nimb, Ida Flörke, Agnes Aggergaard Mikkelsen,
Nathalie Hau Norman & Nicolai Hartvig Sørensen

AI USED TO DETECT PROBLEMATIC CONTENT IN CORPUS EXAMPLES

Keywords AI; harmful content; corpus citations; Danish

This paper describes a method based on AI of which the aim is to identify harmful, offensive, or stereotypical content in corpus citations used to exemplify word meanings in the first, printed version of Den Danske Ordbog (the Danish Dictionary (DDO) 2003-2005). Today the dictionary is published online and is continuously being revised. First, we test to which degree OpenAI's Moderation Model can identify 234 problematic citations that have already been removed, and we compare with a dataset consisting of the neutral citations inserted to replace them. 80% of the problematic citations are identified, and the model is then applied on 71,000 citations from unrevised entries. 6.3% are identified as harmful, the vast majority due to violent content, a type that has been overlooked in previous DDO revisions. A random sample is evaluated, and in the case of violent content, the precision is only 10%, mainly due to many misinterpreted citations. In the case of hateful or/and sexual content, it is 33%. The method results in around 700 validated examples that will be used as input in the editorial process. Two lexicographers must agree whether to remove the identified examples, which will then be replaced by unproblematic ones.

Contact information

Sanni Nimb
Det Danske Sprog- og Litteraturselskab (DSL)
sn@dsl.dk

Ida Flörke
Det Danske Sprog- og Litteraturselskab (DSL)
if@dsl.dk

Agnes Aggergaard Mikkelsen
Det Danske Sprog- og Litteraturselskab (DSL)
aam@dsl.dk

Nathalie Hau Norman
University of Copenhagen
naha@hum.ku.dk

Nicolai Hartvig Sørensen
Det Danske Sprog- og Litteraturselskab (DSL)
nhs@dsl.dk

Krzysztof Nowak, Iwona Krawczyk & Jagoda Marszałek

DIALEXPOL

LLM-Assisted Sense Assignment for a Diachronic Dictionary of Latin

Keywords Word sense disambiguation; Neo-Latin; LLMs; historical lexicography

The DiaLexPoL project is building a diachronic lexical database of Neo-Latin (ca. 1550–1800) on the sense inventory of the Dictionary of Medieval Latin in Poland (DMLP). Its principal compilation bottleneck is sense assignment: referring early-modern attestations to an inventory designed for medieval Latin. We test whether deployable LLMs – cheap commercial mid-tier and self-hostable open-weight models, rather than frontier systems a chronically underfunded project cannot reproduce – can draft this step in a human-in-the-loop workflow. Two experiments test this: reproducing DMLP's own sense assignments on its quotations, and assigning Neo-Latin concordance lines to the DMLP inventory against independent human gold. Self-hostable open-weight models give the strongest drafts, outperforming the single commercial mid-tier route at negligible cost, and the best corpus configuration matches human inter-annotator agreement overall, though not on every lemma. But gains over a most-frequent-sense baseline – always assigning a lemma's commonest sense – are modest and concentrate on genuinely polysemous lemmas, where editors most need help, while on dominant-sense lemmas the models over-discriminate and should defer to that baseline. We treat LLM sense-assignment as a targeted post-editing aid, not an autonomous step, and release the prompts with both labelled datasets.

Contact information

Krzysztof Nowak
Institute of Polish Language (Polish Academy of Sciences)
krzysztof.nowak@ijppan.pl

Iwona Krawczyk
Institute of Polish Language (Polish Academy of Sciences)
iwona.krawczyk@ijp.pan.pl

Jagoda Marszałek
Institute of Polish Language (Polish Academy of Sciences)
jagoda.marszalek@ijp.pan.pl

Krzysztof Nowak & Dorota Mika

POLILEX: MODELLING HETEROGENEOUS LEXICAL RESOURCES

Lessons from the Dictionary of Polish Dialects

Keywords Linguistic linked open data; OntoLex-Lemon; retro-digitised dictionaries; lexicographic resource integration; Polish dialectology

The Institute of the Polish Language (Polish Academy of Sciences) maintains a body of dictionaries and card-file archives assembled over more than a century, whose differences in structure, editorial convention, and descriptive metalanguage have long kept them difficult to search together. This paper reports on PoliLex, an effort to bring such material under a common, machine-queryable representation without erasing the editorial character of each source. We describe an encoding and publication workflow that moves corrected dictionary text through TEI into an OntoLex-Lemon Linked Data representation, and a layered project ontology whose load-bearing decision is to keep a dictionary, as a published text, separate from the language system it records. We illustrate the approach with the Dictionary of Polish Dialects, the first resource taken through the full pipeline, and show how the same encoded source is served both as browsable full text and as a SPARQL-queryable graph. We then examine what such harmonisation is designed to achieve across resources, what it achieves today on a single resource, and where it stops — notably at cross-dictionary sense alignment.

Contact information

Krzysztof Nowak
Institute of Polish Language (Polish Academy of Sciences)
krzysztof.nowak@ijppan.pl

Dorota Mika
Institute of Polish Language (Polish Academy of Sciences)
dorota.mika@ijppan.pl

Emma Josefin Ölander Aadland

A CRITICAL LEXICOGRAPHICAL DISCOURSE ANALYSIS OF NEW ADJECTIVAL ENTRIES IN THE STANDARD DICTIONARIES FOR NORWEGIAN

Keywords Gender and language; gender stereotypes; lemma inclusion; lemma selection; critical lexicography; Critical Lexicographical Discourse Studies (CLDS); gender critical corpus analysis; socio-semantic categorization; concordance analysis

The study presented in the article examines the lemma selection of newly added adjectives in the Norwegian standard dictionaries *Bokmålsordboka* (BOB) and *Nynorskordboka* (NOB). Anchored in *Critical Lexicographical Discourse Studies* (CLDS), the study investigates whether the new adjectives and their sentence examples in the dictionary-entries reflect potential gender-related biases or asymmetries. Methodologically, the new adjectives are classified and analysed using a socio-semantic category system, combined with concordance analysis of sentence examples to examine patterns of gender representation. The findings show largely comparable patterns across the two dictionaries. There is a clear tendency for new adjectives to be associated with masculine-labelled subcategories versus feminine ones. Furthermore, the sentence examples in the dictionary-entries for the new adjectives are predominantly gender-neutral, though a weak tendency towards male overrepresentation is observed, particularly in NOB. Interpreted through a CLDS perspective, the study demonstrates how lexicographic practices related to lemma inclusion may subtly reproduce or cautiously negotiate gendered representations, offering insights relevant to ongoing revision work and lexicographic methodology.

Contact information

Emma Josefin Ölander Aadland
University of Bergen Library
emma.aadland@uib.no

Benedikt Perak, Dragana Špica & Irena Srdanović

A GRAPH-BASED LEXICOGRAPHIC WORKFLOW FOR JAPANESE ADJECTIVES AND THEIR LEXICAL RELATIONS IN AN AI TUTOR

Keywords Japanese adjectives; lexical relations; POS variants; lexical graph; Neo4j; CALL; AI-assisted lexicography

This paper presents a graph-based lexicographic workflow for modelling Japanese adjective lexical relations, integrating reading-aware lexeme identity and POS-variant disambiguation within an AI-assisted relation-building pipeline supported by graph-structural validation. The approach addresses two challenges in Japanese lexicography: kanji homography, where a single orthographic form corresponds to multiple readings, and grammatical ambiguity arising from the adjectival–nominal continuum of na-adjectives. Word nodes are keyed to (standard_orthography, reading_hiragana) pairs, ensuring homographs remain distinct. POS-variant nodes linked via POS_VARIANT_OF edges capture grammatical ambiguity while constraining each node to appropriate relation types. Nine controlled relation types are organised into three families: synonymy (SYNONYM, NEAR_SYNONYM), antonymy (GRADABLE_ANTONYM, COMPLEMENTARY_ANTONYM, SCALAR_INTENSITY), and lexical variation (REGISTER_VARIANT, DOMAIN_VARIANT, MORPHOLOGICAL_VARIANT, ORTHOGRAPHIC_VARIANT). The pipeline imports 1,363 provable edges from JMdict, converts these into deterministic candidates, and employs GPT-4.1-mini and GPT-5.4-mini for structured classification. Every candidate passes through five validation gates: self-loop prevention, duplicate detection, reading-aware resolution, POS-relation consistency, and collocation detection. A pilot on 400 adjectives yielded 4,926 proposals; 283 (79%) were accepted by three expert linguists (mean 4.83/5). Validation achieved 100% recall on identity, POS, and reading checks. The 91 accepted LEXICAL_RELATION edges drive exercise templates in an AI Japanese tutor.

Contact information

Benedikt Perak
University of Rijeka
bperak@uniri.hr

Dragana Špica
Juraj Dobrila University of Pula
dspica@unipu.hr

Irena Srdanović
Juraj Dobrila University of Pula
isrdanovic@unipu.hr

Linda Prossliner

AI-SUPPORTED PHRASEOGRAPHY FOR YOUNG LEARNERS

Designing Definitions and Illustrations through AI – Opportunities and Challenges

Keywords Phraseological learner's dictionary; L2 phraseography; phraseodidactics; digital lexicography; children's dictionaries; AI-assisted definition generation; AI-generated illustrations; generative AI

This paper presents a case study exploring the potential of Large Language Models for learner-oriented phraseography through the generation of COBUILD-style Full-Sentence Definitions (FSDs) and AI-generated dual illustrations representing both the literal and phraseological meanings of idioms. The study is situated within *Phrasolino*, an electronic Italian–German phraseological learner's dictionary for German-speaking children aged 7–11 learning Italian as an L2. A corpus-based dataset of 100 Italian idioms containing the body-part noun *testa* ('head') was compiled from the Italian Web Corpus itTenTen20 and complementary lexicographic resources. ChatGPT-5.5 was prompted to generate child-oriented FSDs and corresponding dual illustrations. The outputs were evaluated qualitatively with regard to semantic accuracy, phraseological adequacy and pedagogical suitability. The findings show that ChatGPT successfully reproduces the formal conventions of COBUILD-style definitions, although recurrent difficulties emerge in the semantic simplification and age-appropriate adaptation of phraseological meaning. The image-generation analysis further reveals both the potential and the limitations of AI-assisted multimodal phraseography, particularly regarding semantic disambiguation, literal interpretation and representational bias. Overall, the study highlights the opportunities of generative AI for learner-oriented lexicography while emphasizing the continuing necessity of lexicographic supervision and post-editing.

Contact information

Linda Prossliner
University of Innsbruck
Linda.Prossliner@uibk.ac.at

Laura Rebosio

OSTENSIVE AIDS IN LEARNER'S DICTIONARIES

A Diachronic, Comparative Analysis

Keywords Illustrations; English learners dictionaries; prototypes; cultural connotations; polysemy

This paper proposes a diachronic analysis of the use of ostensive aids in 35 print editions and the current online versions of six major English monolingual learners' dictionaries (Oxford, Longman, Collins, Cambridge, Macmillan, Merriam-Webster). Alongside macrostructural features of illustrations (density, types, placement, labeling), I analyze their content, performance, and consistency with respect to three questions: (1) whether principles for selecting lemmas to illustrate are stated or inferable; (2) how polysemy and semantic fields are represented; and (3) how culture-specific concepts are handled. The findings show a persistent lack of explicit, operational selection criteria or principles (with LDOCE2 as a partial exception), a substantial reduction in the use of illustrations in the microstructure in print editions, and an unsystematic implementation of ostensive aids in online editions despite the absence of space constraints. The treatment of polysemy, prototypes, and culture-bound items remains uneven. The present analysis is a preliminary step toward outlining a theoretically grounded methodology for the implementation of ostensive aids in the *Phrase-based Active Dictionary* (see Giacomini, 2025; DiMuccio-Failla, 2025).

References

DiMuccio-Failla, P. (2025). A theory for a usage-based cognitive lexicography. In L. Giacomini & V. Piunno (Eds.), *Patterns of meaning in lexicography and lexicology* (pp. 19–90). De Gruyter. https://doi.org/10.1515/9783111545943

Giacomini, L. (2025). *The directionality of (inter)subjectification in the English noun phrase: Pathways of change*. De Gruyter. https://doi.org/10.1515/9783110338751

Contact information

Laura Rebosio
Universität Innsbruck
laura.rebosio@uibk.ac.at

Lydia Risberg, Kristina Koppel, Margit Langemets,
Hanna Maask, Esta Prangel, Maria Tuulik & Silver Vapper

ARE BEARS MORE OFFENSIVE THAN DOGS WHEN DRUNK?

Using LLMs to Systematize the Labelling of Estonian Synonyms

Keywords Large language models; synonyms; offensive language; register labels; benchmarking; Estonian

This paper investigates the potential of large language models (LLMs) as assistants with corpus analysis in lexicography, focusing on the assignment of register labels in the *EKI Combined Dictionary* (CombiDic). Inconsistencies in register labels become evident across synonym sets in CombiDic, which is the result of different lexicographers working on words at different times. We examine how LLMs from Anthropic, Google, and OpenAI handle the challenge of distinguishing between offensive and colloquial usage, which pose a challenge in lexicographic practice. Using a benchmark of 297 words evaluated by five native Estonian-speaking annotators and three LLMs queried via API, we find that LLMs are quite consistent tools for register categorisation and that corpus context significantly improves their performance. Depending on the LLM, 80–89% of suggestions were deemed acceptable for lexicographic use, although since LLMs may be trained to detect offensive language, they tended to select the OFFENSIVE label more often than human annotators. Nevertheless, Gemini 3.1 Pro performed best overall. The experiment already enabled corrections to CombiDic, demonstrating practical value. While full systematisation has not yet been achieved – the broader goal is labels grounded in usage data – LLMs represent a promising and time-efficient aid in semi-automatic lexicographic workflows.

Contact information

Lydia Risberg
Estonian Language Institute
lydia.risberg@gmail.com

Kristina Koppel
Estonian Language Institute
kristina.koppel@eki.ee

Margit Langemets
Estonian Language Institute
margit.langemets@eki.ee

Hanna Maask
Estonian Language Institute
hanna.maask@eki.ee

Esta Prangel
Estonian Language Institute
esta.prangel@eki.ee

Maria Tuulik
Estonian Language Institute
maria.tuulik@eki.ee

Silver Vapper
Estonian Language Institute
silver.vapper@eki.ee

Ines Röhrer & Manuel Raaf

REFLECTIONS ON THE SEMANTIC UNDERSTANDING OF GENERATIVE AI

From LLM Experiments to Lexicographical Tools

Keywords Lexicography; large language models; semantic classification; editorial workflows

This article presents an application evaluation of the use of large language models (LLMs) for the semantic classification of lexicographical content in three German dialect dictionary projects. Building on earlier experiments with LLM-based semantic classification that showed hit rates of over 80%, the study investigates whether generative AI can support the assignment of meanings to controlled semantic categories while preserving editorial oversight. Evaluation data from user ratings, surveys, and interviews indicate that the tool produces useful classifications in a majority of cases, reduces manual research effort, and contributes to greater consistency in semantic annotation. At the same time, the results reveal recurring limitations, especially in cases involving semantic granularity, context-dependent meanings, and definitions where modifiers are overemphasized. The paper argues that LLMs are most valuable not as replacements for expert lexicographical work, but as components of hybrid workflows that combine automated suggestion with human validation.

Contact information

Ines Röhrer
Bavarian Academy of Sciences and Humanities
ines.roehrer@badw.de

Manuel Raaf
Bavarian Academy of Sciences and Humanities
raaf@badw.de

Antonio San Martín & Catherine Trekker

INTRODUCING HUMAN-CENTEREDNESS IN AI-ASSISTED LEXICOGRAPHY

Keywords Human-centered AI; generative AI; AI-assisted lexicography

This paper proposes a human-centered artificial intelligence (HCAI) framework for AI-assisted lexicography. While generative AI offers significant opportunities to enhance lexicographic work, it also raises concerns regarding the future role of lexicographers and the preservation of linguistic and cultural diversity. Drawing on HCAI principles and previous applications in other language professions, the paper identifies four interrelated dimensions through which AI integration in lexicography can be understood and critically examined: the augmented lexicographer, the sociotechnical context of AI integration, bias, and the design of AI-powered lexicographic tools. The framework argues that AI should augment rather than replace lexicographers, combining high levels of automation with meaningful human control. It further emphasizes the importance of preserving professional agency, mitigating AI-generated biases, and designing tools around the needs of lexicographers. By doing so, the paper provides a foundation for future research and the beneficial integration of AI into lexicographic workflows

Contact information

Antonio San Martín
University of Quebec in Trois-Rivières
antonio.san.martin.pizarro@uqtr.ca

Catherine Trekker
University of Quebec in Trois-Rivières
catherine.trekker-seguin@uqtr.ca

Saghar Sharifi & Abolfazl Alamdar

TOWARDS A HANDBOOK OF PERSIAN LEXICOGRAPHY

A Lexicographical-Functions Analysis of 20th–21st Century Dictionaries

Keywords Persian dictionaries; function theory; Handbook of Lexicography; lexicographical functions

This article examines four major Persian dictionaries—*Dehkhoda* (1940), *Mo'in* (1972), *Sokhan* (2003), and *Aryanpur* (2009)—through Tarp's Function Theory of Lexicography (2008). Using a descriptive-analytical mixed-methods design, the study evaluates the dictionaries in terms of four lexicographical functions: text reception, text production, translation, and knowledge acquisition. The analysis is based on 12 lexical and grammatical categories, with three sampled lexical items from each category, yielding 144 dictionary-entry observations across the dataset. The findings reveal clear functional asymmetries. *Dehkhoda* performs most strongly in knowledge acquisition and text reception; *Mo'in* preserves much of this interpretive orientation in a more compact form; *Sokhan* demonstrates the strongest monolingual performance in text reception and contemporary communicative usability; and *Aryanpur*, as a bilingual dictionary, is highly effective for translation but very weak in knowledge acquisition. Overall, the study shows that modern Persian lexicography has moved gradually toward greater user orientation, though without achieving balanced multifunctionality. It concludes by arguing for stronger dialogue between Persian lexicography and international lexicographical scholarship and recommends the compilation of a *Handbook of Persian Lexicography*

References

Anvari, H. (Ed.) (2003). *Farhang-e bozorg-e Sokhan [The Great Sokhan Dictionary*]. Sokhan Publications.

Aryanpur Kashani, M. & Assi, M. (2009). *The Aryanpur Progressive Persian-English Dictionary*. Jahan Rayaneh Computer World Co.

Dehkhoda, A. (1940). *Loḡatnāme-ye Dehkhoda* [*Dehkhoda Dictionary*]. Tehran University Press.

Mo'in, M. (1972). *Farhang-e fārsi-ye Mo'in* [*Moin Persian dictionary*]. Amir Kabir.

Tarp, S. (2008). *Lexicography in the borderland between knowledge and non-knowledge: General lexicographical theory with particular focus on learner's lexicography*. Max Niemeyer Verlag.

Contact information

Saghar Sharifi
Department of Teaching English and Translation, Ka.C., Islamic Azad University, Karaj
saghar_sharifi@yahoo.com

Abolfazl Alamdar
Allameh Tabataba'i University, Tehran
abalfazl.alamdar@yahoo.com

Jóhannes B. Sigtryggsson

THE EFFECTS OF PURISM ON A NINETEENTH-CENTURY ICELANDIC BILINGUAL DICTIONARY

Keywords Purism; historical dictionaries; Icelandic; loanwords

The Icelandic linguist and philologist Konráð Gíslason (1808–1891) was the author of an influential Danish–Icelandic dictionary published in 1851, the first bilingual dictionary of its kind. The significance of this work lay less in its impact on lexicography than in its role in the history of linguistic purism in Iceland. One of its most distinctive features was the near absence of loanwords used as Icelandic equivalents for Danish headwords, despite the fact that such loanwords were common in both spoken and written Icelandic at the time. Gíslason adopted a strategy of drawing upon, and at times reviving, archaic vocabulary attested in medieval texts as translations of the lemmas. Where no suitable archaic term could be identified, he frequently paraphrased the meaning of the Danish entries in explanatory phrases rather than resorting to loanwords. Although the vigorous coinage of neologisms was a hallmark of nineteenth-century linguistic purism, Gíslason's dictionary contains relatively few newly coined terms.

Contact information

Jóhannes B. Sigtryggsson
Árni Magnússon Institute in Icelandic Studies
j.sigtryggs@gmail.com

Hana Skoumalová

EXPANDING THE MWE DATABASE WITH VARIANTS

Keywords Corpus annotation; phraseology; database expansions

This paper presents recent extensions to the Czech Lemur MWE database, focusing on the systematic generation of syntactic transformations and morphological derivations to improve phraseological corpus annotation. The database currently contains over 29,500 entries, each enriched with information on positional variability, optional components, and synonymy, enabling the mwe_tagger to identify a wide range of attested variants. We describe a rule-based approach to generating periphrastic passives, reflexive passives, and deagentive constructions, as well as selected derivational variants supported by MorfFlex dictionary and DeriNet tool. While simple derivations (e.g., nominalizations or participial adjectives) can be safely automated, more complex cases require manual validation due to semantic drift or structural constraints. To evaluate the impact of these expansions, we compare annotations in SYNv14 and two versions of the Frap corpus. Although transformed or derived MWEs remain relatively infrequent (2–3.5% of all MWE tokens), their inclusion increases annotation coverage and reveals additional phraseological behavior. These results demonstrate that controlled automatic expansion is both feasible and beneficial for large-scale MWE annotation.

Contact information

Hana Skoumalová
Charles University
hana.skoumalova@ff.cuni.cz

Deepak Solanki

USER-ORIENTED SPECIALIZED LEXICOGRAPHY

Corpus-Driven Models for Targeted Dictionary Design

Keywords Specialized lexicography; corpus-driven lexicographic models; targeted dictionary design; recommender systems; Śabdakalpadruma; Sanskrit lexicography; digital humanities

Lexicographic practice has often balanced comprehensive coverage with systematic ordering, privileging alphabetical arrangement and encyclopedic completeness over clearly defined user needs. This paper argues that user-oriented specialized dictionaries, supported by corpus-driven and AI-augmented methods, can better serve users who require targeted lexical knowledge rather than exhaustive inventories. The study uses Śabdakalpadruma as a Sanskrit lexicographic case study and reports its conversion into a structured digital corpus of 42,824 headword entries through OCR, post-OCR cleaning, Excel inspection, SQLite structuring, JSON representation, and searchable HTML implementation. Building on this corpus, the paper develops a pilot rule-based recommender model for ritual vocabulary. Twenty ritual query terms were searched in headword and definition fields, and the extracted entries were automatically annotated through semantic tagging, co-occurrence detection, relation scoring, confidence labelling, and relevance classification. A stratified validation sample of 100 entries, with five entries from each query term, was manually evaluated. The results show that 67% of annotations were correct, 19% usable with correction, and 14% wrong. The study demonstrates a staged pathway from digitized lexicographic corpus to rule-based recommendation, human-validated training data, and future AI-assisted semantic prediction.

Contact information

Deepak Solanki
Jawaharlal Nehru University, New Delhi
deepsoul.solanki@gmail.com

Michael Spagnol, John J. Camilleri & Dwayne Ellul

FROM *ĠAMMA* TO *ĠABRA*

Four Centuries of Maltese Lexicography

Keywords Maltese lexicography; digital resources; dictionary structures; lemmatisation

This paper presents a historical and typological overview of Maltese lexicography, tracing the development of lexical resources from the earliest manuscript dictionaries to recent digital initiatives. It surveys major multilingual, bilingual, monolingual and specialised dictionaries while discussing the methodologies underlying Maltese lexicography over the years. Emphasis is given to the interaction between the hybrid lexical structure of Maltese and lexicographic practice, including issues of lemmatisation, orthography and dictionary structure. The paper also discusses recent digital initiatives, particularly Ġabra and Dizzjunarju.mt, which illustrate the growing role of computational methods in Maltese lexicography.

Contact information

Michael Spagnol
L-Università ta' Malta
michael.spagnol@um.edu.mt

John J. Camilleri
Chalmers University of Technology and the University of Gothenburg
john.j.camilleri@chalmers.se

Dwayne Ellul
L-Università ta' Malta
dwayne.ellul@um.edu.mt

Vasyl Starko & Andriy Rysin

LARGE-SCALE SEMIAUTOMATIC DETECTION OF NEW AND UNREGISTERED WORDS IN MODERN UKRAINIAN

Keywords Ukrainian dictionary; neologism; neologism detection methodology; Ukrainian corpus; unregistered lexis; Ukrainian lemmatizer; Ukrainian POS tagger

We have created and implemented a workflow for large-scale semiautomatic detection of new and previously unregistered words in Modern Ukrainian. Crucially, it involves a diverse corpus of Ukrainian, a large electronic morphological dictionary of Ukrainian, and a specially crafted NLP toolkit. The key methodological innovation lies in the dual use of the dictionary. On the one hand, it serves as a resource for lemmatizing and morphological tagging the corpus; on the other, it is cyclically enriched with new words detected in the corpus. The corpus is dynamic, as it is regularly expanded through the addition of a variety of texts from the 19th century to the present day. The dictionary, currently containing 442,000 lemmas, is dynamically reused with the help of a specially crafted NLP toolkit as an exclusion source (Janssen 2009) against the growing corpus. The paper provides a detailed step-by-step description of the procedure. This semiautomatic pipeline has yielded thousands of new and unregistered Ukrainian words, laying the foundation for the systematic, large-scale exploration of the Modern Ukrainian lexicon.

Contact information

Vasyl Starko
Ukrainian Catholic University
v.starko@ucu.edu.ua

Andriy Rysin
Independent researcher
arysin@gmail.com

Nika Stefan, Johan van der Zwaag & Hindrik Sijens

BEYOND THE SINGLE STANDARD

Addressing Variation and User Needs in an Online Frisian Dictionary

Keywords Labelling; online dictionaries; minority language; language variation; standardization; description/prescription; lexicographic practice

This paper examines the challenges of representing linguistic variation in Frisian lexicography, focusing on the Dutch-Frisian online dictionary. While existing lexicographic resources already accommodate variation to some extent, discrepancies remain between dictionary representation and actual language use. Survey data and usage examples reveal complex patterns in which standard forms coexist with contact-induced variants, and in some cases no fully equivalent Frisian alternatives exist, leading speakers to adopt Dutch forms in Frisian discourse. These findings highlight a tension between normative guidance and descriptive adequacy, particularly in contexts where speakers vary in competence and where written Frisian differs from spoken usage. The current label system, although informative, does not always provide sufficiently systematic or user-oriented support. Building on insights from user-oriented lexicographic theory, this paper proposes a more flexible model in which dictionary information is accessed through multiple modes. These include a normative mode offering clear guidance, a comparative mode making variation explicit, and a learning mode tailored to users with less experience. By structuring variation in a transparent and user-sensitive way, the proposed approach aims to align lexicographic practice more closely with actual language use while supporting both standardization and linguistic diversity.

Contact information

Nika Stefan
Fryske Akademy
nstefan@fryske-akademy.nl

Johan van der Zwaag
Fryske Akademy
jvanderzwaag@fryske-akademy.nl

Hindrik Sijens
Fryske Akademy
hsijens@fryske-akademy.nl

Petra Storjohann

THE INTERPLAY BETWEEN INDIVIDUAL NEW LEXICAL ITEMS AND MULTI-LAYERED DISCOURSE REPRESENTATIONS OF NEOLOGISMS

Keywords Controversial neologisms; CADS; discourse lexicography; discourse entries; IDS *Neo*$^{2020+}$

This paper examines how discourse lexicography can be extended to the study of contemporary neologisms, particularly those emerging from major societal crises such as COVID-19, climate change, energy shortages, migration, inflation and the war in Ukraine. Traditional discourse lexicography has mainly focused on historically controversial terms and has typically relied on selected textual sources rather than large-scale corpora. In contrast, the paper argues for a corpus-assisted discourse-analytic approach that systematically captures the semantic, evaluative and ideological dimensions of newly coined and borrowed lexical items. Particular attention is paid to features such as semantic prosody, affective meaning, stance-taking, politicisation and power relations, which often become entrenched in neologisms and shape their use in public discourse. The paper introduces the German neologism resource IDS *Neo*$^{2020+}$, which combines two complementary levels of consultation: detailed lexical entries for individual neologisms and broader discourse or cluster entries that trace thematic developments and discourse networks over time. Both entry types are interconnected. This dual structure allows users to explore both the meanings of specific new words/expressions and their wider discursive functions and relations with other new terms within public communication. The paper also outlines the corpus-linguistic and computational methods used in the project, including word embeddings.

Contact information

Petra Storjohann
IDS Mannheim
storjohann@ids-mannheim.de

Silga Sviķe

A LEXICOGRAPHICAL ANALYSIS OF NEOLOGISM DICTIONARIES BASED ON THE MULTILINGUAL DICTIONARY OF NEW WORDS (2010–2024)

Keywords Neologism; new word; dictionary of new words; lexicography

The paper examines the structure and treatment of neologisms in eight editions of the Multilingual Dictionary of New Words and focuses specifically on the representation of Latvian material within this multilingual European resource. The study is based on mega-, macro-, and microstructure of dictionaries and gives a wide dictionary-analytical basis within Latvian neologism dictionaries. The research aim is to evaluate the structural evolution of the Multilingual Dictionary of New Words and examine the representation of Latvian neologisms within it, focusing on the inclusion and stability of Latvian neologisms across its eight editions. The methodological framework of this study is designed around a triangulated approach, combining theoretical analysis, empirical qualitative inquiry, and mixed-methods data analysis. A systematic extraction of Latvian neologisms was conducted across the entirety of the eight-volume dictionary. A dataset of more that hundred specific lexical units in Latvian (and also in Lithuanian) was compiled. The study employs a wide analytical approach: qualitative theoretical source analysis and qualitative analysis of dictionary structure, as well as quantitative and qualitative analysis of excerpted neologisms.

Contact information

Silga Sviķe
Ventspils University of Applied Sciences
silga.svike@gmail.com

Pius ten Hacken

THE THEORETICAL FOUNDATION OF DICTIONARY DEFINITIONS

Keywords Applied science; explanation; definition; theory

The place of theory in lexicography has been the matter of intensive debate. There are different views of what a theory is. In line with most approaches in philosophy of science, it is assumed here that a theory has to provide an explanation. It is argued that lexicography should be considered as an applied science, i.e. the same category as medicine. In the analysis of applied science, the identification of problems and the specification of evaluation criteria for solutions are crucial. In lexicography, problems are the reasons why users consult a dictionary. Problem types are distinguished on the basis of the kind of information they need and the background knowledge they have. For certain problem types, definitions can be used as solutions, but the requirements on the definition depend on the type of problem it is meant to solve. Guidelines issued to lexicographers working on a dictionary are not theories, because they do not explain. They only instruct. The theoretical foundations of a theory should identify criteria for distinguishing problem types.

Contact information

Pius ten Hacken
Universität Innsbruck
pius.ten-hacken@uibk.ac.at

Dušica Terzić, Dejana Mirković-Birtašić & Saša Marjanović

TOWARDS A FRENCH–SERBIAN DICTIONARY OF SPORTS TERMS

Corpus-Based or AI-Assisted Methodology?

Keywords Dictionary project; corpus-based dictionary; AI generated entries; lexicographic processing; French–Serbian dictionary; sporting terminology

Earlier multilingual dictionaries of sports terminology that include French and Serbian were largely based on simple translations from foreign sources, without systematic terminological analysis whereas more recent works do not include French. To be able to address this gap and compile a reference dictionary, we first need to identify the optimal methodology for the compilation of this type of dictionary entries. This paper presents the evaluation of two different possible approaches—corpus-based and AI-assisted generation. First, we selected ten polysemous French sports terms with multiple Serbian equivalents across different contexts and generated dictionary entries using both corpus analysis and AI models. The corpus used in the process is a multilingual parallel corpus ParCoLab containing aligned English, French, and Serbian texts from official sports rulebooks of 14 sports, providing a solid basis for terminological extraction, whereas the AI models evaluated were GPT-5.5, Sonnet 4.6 and Gemini 3.6. The generated entries were then evaluated by human experts for accuracy, consistency, and practical usefulness. The results show that while AI significantly accelerates the compilation process, corpus-based methods remain more reliable. Combining both approaches, thus, proves to be most effective: corpus ensures data quality, while AI supports efficiency. However, final validation by a human terminologist remains essential for refining meanings, standardizing terminology, and ensuring long-term usability.

Contact information

Dušica Terzić
University of Belgrade - Faculty of philology
dusica.terzic@fil.bg.ac.rs

Dejana Mirković-Birtašić
University of Belgrade - Faculty of philology
dejana.mirkovic.birtasic@fil.bg.ac.rs

Saša Marjanović
University of Belgrade - Faculty of philology
sasa.marjanovic@fil.bg.ac.rs

Carole Tiberius, Martin Kroon, Lut Colman & Dalí Chirino

RANKING COLLOCATIONS IN *WOORDCOMBINATIES*

Keywords Collocation measures; Dutch; lexicographic resource

This paper describes efforts to extract frequency information from corpora and to rank the collocates in *Woordcombinaties*, a Dutch integrated lexicographic resource, in a uniform and systematic way, taking various factors such as the size and composition of the corpus, linguistic processing and the mathematical properties of various association measures into account. A small sample of 2036 combinations, divided over 19 verbs and 19 nouns, was used, and their frequencies were extracted from the Corpus Contemporary Dutch (CHN) and various subsets of said corpus, using surface co-occurrence, textual co-occurrence, and syntactic co-occurrence. Results were verified manually. It was found that: 1) corpus size increases coverage; 2) syntactic co-occurrence leads to loss of coverage; 3) combination of a narrow window span and within-sentence context seems to have the highest coverage; 4) separate treatment is needed to deal with multi-word collocates. However, given that surface/textual co-occurrence still produces a substantial amount of false positives, we intend to use syntactic co-occurrence to retrieve the frequency information for the combinations in *Woordcombinaties*, because it has a higher precision and offers the greatest scope for improvement.

Contact information

Carole Tiberius
Instituut voor de Nederlandse Taal
c.p.a.tiberius@hum.leidenuniv.nl

Martin Kroon
Instituut voor de Nederlandse Taal
m.s.kroon@hum.leidenuniv.nl

Lut Colman
Instituut voor de Nederlandse Taal
lut.collman@ivdnt.org

Dalí Chirino
Instituut voor de Nederlandse Taal
dali.chirino@ivdnt.org

Jesús Torres del Rey & M.ª Teresa Fuentes Morán

A HYBRID APPROACH TO WEB ACCESSIBILITY EVALUATION IN DIGITAL LEXICOGRAPHY

Comparing and Integrating Rule-Based Validators and Generative AI

Keywords Web accessibility evaluation; electronic dictionaries; large language models; artificial intelligence; WCAG; screen reader users; blind users; dictionary microstructure; automatic validators

Electronic dictionaries pose particular accessibility challenges for blind screen reader users, yet the field lacks evaluation frameworks that account for the complexity of their information structures. This paper reports on a three-phase study comparing Automated Web Accessibility Evaluation Tools and Large Language Models in evaluating the microstructure of three online dictionary entries from different lexicographical traditions. Phase 1 establishes a baseline through automated WAET scanning and zero-shot LLM auditing. Phase 2 compares their outputs and derives a four-category framework, reframing compliance as a design question. Phase 3 tests two enhanced prompting strategies — literature-augmented and design-oriented. The former uses specific dictionary accessibility literature while the latter combines a structured checklist derived from Phase 1 with persona-based use-type simulation, surfacing a macro-to-micro hierarchy of dictionary-specific barriers. The study argues for an iterative, multi-condition hybrid workflow integrating WAETs, and expert human review for e-dictionaries.

Contact information

Jesús Torres del Rey
Universidad de Salamanca
jtorres@usal.es

M.ª Teresa Fuentes Morán
Universidad de Salamanca
tfuentes@usal.es

Ágoston Tóth

(UG)TPEX

A (Usage-Guided) Typical Example Score

Keywords Example selection; sense alignment; LLM-assisted human lexicography; BERT

The aim of this paper is to explore methods to utilize the linguistic knowledge accumulated in language models for evaluating corpus sentences as potential examples in dictionary entries. We observe the TPEX (Typical Example) measure in three variants in the task of measuring the typicality of a sentence based on how probable the appearance of the headword in the given context is, as per the selected language model. Then we present an extension, the UGTPEX (Usage-Guided Typical Example) measure, in six variants, where the contextual typicality is supplemented by data reflecting the similarity of the headword's contextualized embedding to representations seen in known-good (dictionary-derived, sense-aligned) examples. The combination of the two components, i.e. the contextual typicality and the embedding-based similarity, results in a score that is sensitive to both frequent usage patterns and the usage-categories (sense-categories) established by the lexicographer. We will examine the applicability of these measures in case studies for five headwords using dictionary example sentences and corpus samples. We recommend the proposed scoring procedures primarily for ordering concordance lines, in combination with other ranking methods, as part of the lexicographic workflow.

Contact information

Ágoston Tóth
Department of English Linguistics, University of Debrecen
toth.agoston@arts.unideb.hu

Ágoston Tóth & Esra Abdelzaher

GPT AS A SENSE-TAGGER FOR SELECTING DICTIONARY EXAMPLES

Keywords Dictionary examples; GPT; sense alignment; sense delineation

Aligning a target word in a sentence with a dictionary sense is a challenging task, given the dynamicity of meaning (in context) and the fixed discrete senses in dictionaries, among other reasons. However, it is still a practical necessity for a lexicographer to select example sentences that best illustrate dictionary senses and, hence, find a way to address this challenge. Further challenges appeared after the established use of large corpora as sources of lexicographic evidence and the need to update dictionary entries with additional illustrative examples. The present study tested the usability of an LLM as a facilitator in sense aligning and example selecting tasks. The role of the model was limited to assign a score that reflects the match between (corpus-based and dictionary-cited) examples and existing dictionary senses under several experimental conditions. Quantitative and qualitative analyses showed the usefulness of the model in selecting candidate senses that best represent a dictionary sense, detecting misalignment cases in dictionaries, excluding less representative examples despite their high GDEX scores and identifying highly overlapping senses (especially in the category of adjectives) that are likely to be puzzling to the users.

Contact information

Ágoston Tóth
Department of English Linguistics, University of Debrecen
toth.agoston@arts.unideb.hu

Esra Abdelzaher
Department of English Linguistics, University of Debrecen
esra.abdelzaher@gmail.com

Lars Trap-Jensen & Henrik Lorentzen

REVISING THE USAGE-LABELLING SYSTEM OF THE DANISH DICTIONARY

Keywords Usage labels; marked language; discriminatory language; dictionary revision; sociolinguistics; Danish lexicography; identity terms; corpus lexicography

This article discusses the ongoing revision of the usage-labelling system in *Den Danske Ordbog* (DDO, 'The Danish Dictionary') in response to increasing public attention on discriminatory and otherwise socially marked language. In recent years, the dictionary has received growing criticism concerning the treatment of identity-related terms, revealing limitations in the existing inventory of stylistic and evaluative labels. The article examines the lexicographic challenges involved in balancing descriptive dictionary practice with users' expectations for guidance and contextualisation. To address these issues, DDO is developing a revised system based on a closed inventory of evaluative labels, optional modifiers, and controlled combinations. The new model is supplemented by graphical warning symbols and expandable usage notes inspired by practices in dictionaries such as Merriam-Webster and Duden. The proposed revision seeks to establish a compromise between systematic description, user guidance, and the scientific legitimacy of descriptive lexicography.

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

Thordis Ulfarsdottir & Ellert Thor Johannsson

THE REVERSAL OF A BILINGUAL DICTIONARY USING AI

Keywords Icelandic; bilingual lexicography; reversal; LLM

This paper presents a pilot study exploring the use of artificial intelligence in reversing a bilingual dictionary. The Modern Icelandic-English Dictionary is a freely available online resource. The aim was to transform this Icelandic-English dictionary into a functional English-Icelandic dictionary while minimizing manual labour. Using the DANTE English lexical database as a reference, the study focused on English words beginning with the letter m. The process was divided into five phases. First, SQL queries were used to generate an initial reverse listing. Second, the Gemini large language model was employed to process the output and combine it with data from the DANTE list. The resulting dataset contained approximately 4,700 potential entries, of which 1,702 lacked an Icelandic equivalent. In the third phase, Gemini supplied missing translations, while in the fourth phase it performed sense disambiguation for polysemous entries. The fifth phase involved post-processing, including the elimination of unnecessary entries and extensive proofreading. The findings suggest that AI can successfully handle three of the five phases, thereby significantly reducing the workload traditionally carried out by human lexicographers. However, substantial post-processing still requires human expertise, although AI may also be used to facilitate parts of that work.

Contact information

Thordis Ulfarsdottir
The Árni Magnússon Institute for Icelandic Studies
disa@hi.is

Ellert Thor Johannsson
The Árni Magnússon Institute for Icelandic Studies
etj@hi.is

Anna Vacalopoulou

A DEFINITION WALKS INTO A DICTIONARY

The Lexicography of Humor

Keywords Lexicography; dictionaries; abstract concept definition; microstructural analysis; humor studies

Different disciplines approach the definition of humor from different angles: philosophy through theoretical frameworks that explain why people find things funny, psychology through behavioral and dispositional dimensions that describe how humor is experienced and expressed. This paper examines the treatment of humor in terms of its lexicographic definition, drawing on entries from six Greek dictionaries currently in circulation. Through systematic microstructural analysis, definitional components are identified and categorized, revealing significant variation in how dictionaries conceptualize humor. Definitions are broken down into components that reveal lexicographers' attitudes towards humor's nature (what it is), manifestation (how it appears), subject (what its target is), features (what characterizes it), and function (what its goal is). Results demonstrate that while lexicographers employ diverse definitional strategies, ranging from analytical to functional approaches, they align on certain core elements, particularly humor's cognitive dimension and social function. This research proposes a systematic methodology for analyzing how dictionaries handle abstract, culturally embedded concepts, providing a replicable framework applicable to other affective, cognitive, or socially constructed concepts such as irony, empathy, and creativity.

Contact information

Anna Vacalopoulou
Institute for Language and Speech Processing / Athena RC
avacalop@athenarc.gr

Ene Vainik, Heete Sahkai, Ahto Kiil,
Jelena Kallas & Geda Paulsen

SEMANTIC CLUSTERING AND CATEGORISATION OF CONSTRUCTIONAL COLLEXEMES USING LLMS

Keywords Constructicography; lexicography; computational semantics; colloprofile; Estonian

This paper investigates whether large language models can perform semantic clustering and categorisation of constructional collexemes to support the analysis of constructional meaning and the organisation of collexemes within constructicon entries. As a case study, we examine the collexemes of the Estonian Nominal Quantifier Construction, identified from lexicographic and corpus data. Using OpenAI's ChatGPT-5.4, we conducted an informed free sorting task and a closed sorting task with five input types, consisting either in bare lemmas or lemmas accompanied by different types of context: corpus phrases, corpus sentences, dictionary definitions, and dictionary examples. Model outputs were evaluated against a human-created gold standard using Adjusted Mutual Information, Adjusted Rand Index, and a label quality rating scheme. A scaled closed sorting experiment was also conducted. The results of the free sorting tasks approached human agreement levels, with dictionary definitions yielding the most similar clustering and corpus sentence input the most acceptable labels. The results of the closed sorting task and the scaling experiment demonstrated that a bare list of lemmas provided sufficient input for assigning collexemes to predefined semantic categories.

Contact information

Ene Vainik
Institute of the Estonian Language
Ene.Vainik@eki.ee

Heete Sahkai
Institute of the Estonian Language
Heete.Sahkai@eki.ee

Ahto Kiil
Institute of the Estonian Language
ahto.kiil@eki.ee

Jelena Kallas
Institute of the Estonian Language
jelena.kallas@eki.ee

Geda Paulsen
Institute of the Estonian Language
geda.paulsen@eki.ee

Carlos Valcárcel Riveiro, María Teresa Sanmarco Bande,
Iván Arias-Arias & María José Domínguez Vázquez

CAN LEXICOGRAPHIC ENTRIES EMBED DIDACTIC MODULES?

A Case Study on LLM-Generated Outputs for a Dictionary Interface

Keywords LLMs; PORTLEX; didactic lexicography; valency; prompt engineering

Digital dictionaries can become more than consultation tools when their internal data are sufficiently structured to be reused and transformed into embedded learning components. This paper evaluates whether large language models (LLMs) can convert PORTLEX entries, represented as reviewed JSON files, into interactive HTML didactic modules integrable into their corresponding dictionary entries. PORTLEX is a multilingual valency-oriented dictionary of the nominal phrase, and its microstructure offers a controlled source for activities on formal links, ordering, complement function and semantic compatibility. The experiment compares a GPT-based custom system and a Gemini-based Gem across three languages, four semantically aligned nouns per language and two model outputs per entry, producing 24 modules and 48 expert evaluations. Results show that both systems generated broadly usable modules and generally preserved the linguistic-valency information encoded in the JSON input. The main limitations did not concern basic HTML functionality, but pedagogical regulation: feedback, correction, scaffolding, distractor quality and ambiguity management. GPT produced more stable outputs, while Gemini showed greater dispersion. Language-based patterns were diagnostic rather than conclusive. The results support treating LLM-generated modules as controlled prototypes within a procedurally replicable workflow, not as ready-made pedagogical products.

Contact information

Carlos Valcárcel Riveiro
Instituto de Investigación Lingua, Universidade de Vigo
carlos.valcarcel.riveiro@gmail.com

María Teresa Sanmarco Bande
Universidade de Santiago de Compostela
mt.sanmarco@usc.es

Iván Arias-Arias
Instituto da Lingua Galega, Universidade de Santiago de Compostela
ivanarias.arias@usc.gal

María José Domínguez Vázquez
Universidade de Santiago de Compostela
majo.dominguez@usc.es

Martina Vokáčová & Anna Marklová

AFFORDANCE NORMS IN CZECH AND ENGLISH

Human–LLM Misalignment in Embodied Lexical Knowledge

Keywords Affordances; large language models; psycholinguistic norms; lexicography; Czech; English; cross-linguistic comparison; embodied meaning; reporting bias

This paper investigates whether large language models can capture the embodied dimension of word meaning by comparing LLM-generated affordance norms with human-produced norms in Czech and English. Affordance norms are systematically collected data on the actions speakers associate with concrete objects. Human data were collected from 30 Czech native speakers in a free-production experiment using 50 nouns. The same stimuli were presented to four LLMs (Claude Haiku 4.5, Gemini 3 Flash, GPT-5, and Llama 4 Maverick) at two temperature settings in both languages; English outputs were additionally compared against human norms from Maxwell et al. (2024). We report four main findings. First, human affordance norms are more variable in Czech. Second, LLMs consistently produce narrower affordance sets than humans, with substantially lower lexical diversity across both languages. Third, lexical overlap between human and LLM affordances is low (shared vocabulary 15–19%; Jaccard index not exceeding 0.25 at $N = 10$) and does not improve substantially with higher temperature or greater language representation in the training data. Fourth, human-only affordances are more embodied and emotionally grounded, while LLM-only affordances cluster around maintenance and support. The findings suggest that human norming data may offer a valuable resource for enriching lexicographic entries.

References

Maxwell, N. P., Huff, M. J., Hajnal, A., Namias, J. M., Blau, J. J. C., Day, B., Marsh, K. L., Meagher, B. R., Shelley-Tremblay, J. F., Thomas, G. F. & Wagman, J. B. (2024). Affordance norms for 2825 concrete nouns. *Behavior Research Methods*, 56, 8480–8491. https://doi.org/10.3758/s13428-024-02486-1

Contact information

Martina Vokáčová
Charles University Prague
martina.vokacova@ff.cuni.cz

Anna Marklová
Charles University Prague
a.marklova@gmail.com

Zongyuan Wang, Valeria Caruso & Anna De Meo

COLLABORATIVE LEXICOGRAPHY FOR HERITAGE LANGUAGE LEARNING

Keywords Heritage language learning; collaborative lexicography; Chinese as a heritage language; vocabulary acquisition; dictionary compilation; lexonomy; literacy development

This paper explores the potential of collaborative lexicography as a pedagogical approach to Chinese heritage language learning. The study focuses on a teaching module based on Lu Xun's short story Village Opera (社戏; Shèxì), combining literary reading, cultural exploration, vocabulary learning, and the compilation of digital dictionary entries with Lexonomy. The framework was first implemented with Chinese heritage learners in Italy and then replicated with two age-matched groups of students in China; one experimental group and one control group. All participants completed cloze tests on 15 target lexical items in literary and non-literary contexts, with a delayed post-test administered three months after the intervention. Results show that the evidence from the heritage group remains inconclusive, mainly due to infrastructural limitations and the absence of a comparable control group in Italy. However, the Chinese experimental group showed substantial vocabulary gains and significantly out-performed the control group, especially with literary-register items. These findings suggest that collaborative dictionary compilation can support vocabulary acquisition and cultural learning and indicate the need for further research with heritage learners under improved experimental conditions.

Contact information

Zongyuan Wang
Università degli Studi di Napoli "l'Orientale"
z.wang1@unior.it

Valeria Caruso
Università degli Studi di Napoli "l'Orientale"
vcaruso@unior.it

Anna De Meo
Università degli Studi di Napoli "l'Orientale"
ademeo@unior.it

Tarrin Wills & Ellert Thor Johannsson

CAN AI-ASSISTED NORMALISATION IMPROVE ACCESS TO HISTORICAL LEXICOGRAPHIC DATA?

Keywords Historical lexicography; language normalisation; large language models

The topic of this article is AI-assisted normalisation and how it can be used in the context of a historical dictionary. We look at data from *A Dictionary of Old Norse Prose*, which is a lexicographic resource on the language of medieval Iceland and Norway. This dictionary has a long tradition of using non-normalised, diplomatic text editions as source for its lexicographic descriptions. As a result the bulk of the citation material remains difficult to access for beginners and non-experts in the language. We tested several LLM's in order to assess whether it would be feasible to provide users with automatic normalisations. The article describes several phases in testing and comparing different LLM's and discussing the types of errors they made. The result show that from the initial experimental phase in 2025 the models have improved significantly and currently are able to provide fairly error-free normalisations. We also mention the option of using modern Icelandic as an intermediate stage in the normalisation process which would facilitate the use of NLP-tools. The conclusion is that LLM-assisted normalisation can support broader access to historical lexicographic data. Such tools combine automated processing with human oversight, allowing historical dictionaries to expand their usability while maintaining scholarly rigour.

Contact information

Tarrin Wills
University of Copenhagen
tarrin@hum.ku.dk

Ellert Thor Johannsson
The Árni Magnússon Institute for Icelandic Studies
etj@hi.is

Shigeru Yamada

PEDAGOGICAL VALUE AND INSIGHTS FROM COMPARING EFL DICTIONARIES

Evidence from *Apparently*, *Be on Speaking Terms*, and *Suffer (From)*

Keywords Dictionary comparison; dictionary instruction; EFL dictionary

Certain aspects of word meaning and usage become visible only when dictionaries are compared directly. This paper examines three cases—*apparently*, *be on speaking terms*, and *suffer (from)*—to show how cross-dictionary analysis reveals patterns that remain hidden in single-dictionary consultation. By comparing definitions, sense divisions, and examples across major EFL dictionaries (CALD, COBUILD, LDOCE, MWALED, and OALD), the study demonstrates the pedagogical value of such comparison.

The analysis of *apparently* shows notable variation: CALD lists three senses, whereas others merge or reduce them. Example sentences further reveal mismatches, such as an OALD example illustrating a sense not explicitly defined there but found in CALD.

The expression *be on speaking terms* illustrates how usage tendencies emerge only through comparison. Dictionaries show that the phrase appears predominantly in the negative, and positive uses typically describe the restoration of communication after conflict.

The verb *suffer* shows distinct transitivity and collocational patterns. Cross dictionary analysis reveals contrasts among intransitive use, *suffer from*, and transitive constructions, clarifying semantic tendencies and demonstrating the value of comparative consultation for more precise lexical understanding.

Overall, dictionary comparison reveals usage patterns invisible in single sources, offering clearer insights for interpretation and learning.

Contact information

Shigeru Yamada
Waseda University
shyamada@waseda.jp

Yelena Yerznkyan & Diana Movsisyan

ENCODING EMOTIVITY IN DICTIONARY DEFINITIONS

A Corpus-Based Study of Verbs of Understanding

Keywords Verbs of understanding; collocational pattern; emotivity; corpus linguistics; lexicography; definitions of words

The study identifies the emotional profiles of verbs of understanding and demonstrates their relevance for corpus-driven lexicography, contributing to corpus-based research on emotivity and offering implications for lexicographic description. The paper examines the interaction between understanding and emotion as reflected in the contextual behaviour of English verbs of understanding and provides empirical evidence for models of emotionally grounded meaning representation, supported by the emotive clusters most frequently associated with understanding. From a lexicographic perspective, the findings substantiate the inclusion of the emotive component in dictionary descriptions of the verbs in question. Thus, the research argues for a more refined approach to emotivity in lexicographic description of verbs of understanding considering that the emotional meaning makes the axes of the semantics of verbs of understanding.

Contact information

Yelena Yerznkyan
Yerevan State University
yerznkyan@ysu.am

Diana Movsisyan
Armenian State University of Economics
movsisyan.diana@gmail.com

Marija Žarković Eriksson

LEGAL LEXICON AND THE E-*DRAE* 1884

A Diachronic Study of Specialised Vocabulary in a General Dictionary Using a Digital Critical Edition

Keywords Legal lexicon; historical lexicography; computational lexicography; *DRAE* 1884; ideology; e-*DRAE* 1884

This paper presents a methodology for the diachronic study of legal lexical units related to criminal law in a nineteenth-century general-language dictionary, the twelfth edition of the Diccionario de la lengua castellana of the by the Real Academia Española (*DRAE* 1884), as analyzed through its critical and digital edition (e-*DRAE* 1884). The study combines a lexicographical examination of how these units are defined, structured and interconnected with their contextualization within the legal, social and ideological framework of nineteenth-century Spain, drawing on the sociocognitive theory of ideology and on the notion of the dictionary as discourse. Selected case studies from the criminal lexicon illustrate continuities and ruptures between the inherited juridical tradition and the emerging liberal order. The paper finally discusses the methodological advantages that the e-*DRAE* 1884 environment offers for research on specialized lexis in historical dictionaries.

Contact information

Marija Žarković Eriksson
Universidad Autónoma de Barcelona
marija.zarkovic@autonoma.cat

PART V

WORKSHOPS

EURALEX 2026

Ilan Kernerman & Kris Heylen

8TH GLOBALEX WORKSHOP ON LEXICOGRAPHY AND NEOLOGY

Keywords Lexicography; neology; human–AI interaction

The Globalex Workshop on Lexicography and Neology (GWLN) has established itself as a leading international forum for researchers and practitioners interested in the relationship between lexical innovation and lexicography. Since its inauguration, GWLN has been organized in conjunction with the annual conferences of the major continental lexicographic associations, including DSNA (2019), Euralex (2020 online, and 2022), Australex (2021), Asialex (2023), Afrilex (2024) and, most recently, eLex (2025)[1]. Throughout these editions, the workshop has brought together lexicographers, scholars of neology, corpus linguists, terminology specialists, language technologists, and tool developers to discuss how new words, word senses and multiword expressions emerge, how they can be detected and evaluated, and how they should ultimately be represented in lexicographic resources.

The eighth edition takes place within the framework of the XXII EURALEX International Congress and is supported by the COST Action *European Network on Lexical Innovation* (ENEOLI, CA22126[2]), featuring a session that is specifically dedicated to the very diverse lexicography-related work coming out of ENEOLI. These collaborations reflect the growing recognition that lexical innovation has become a multidisciplinary research area requiring close cooperation across lexicography, terminology, corpus linguistics, natural language processing, and artificial intelligence.

The central theme of GWLN-8 is the changing role of the lexicographer identifying and documenting neology in the age of artificial intelligence (AI). Lexicography has traditionally been among the first linguistic disciplines to adopt new technologies, from using computers through corpus analysis and statistical methods to electronic dictionaries and linguistic linked data. The emergence of large language models (LLMs), however, represents a qualitative shift, as Generative AI not only provides powerful new tools for supporting lexicographic work, but also increasingly influences the very processes through which lexical innovation emerges and spreads. New words and meanings are no longer created solely within human linguistic communities but are increasingly shaped by AI-assisted writing, machine translation, conversational agents, recommendation algorithms, and other forms of human–machine interaction. Consequently, lexicographers are faced with the dual challenge of documenting an evolving lexical landscape while critically assessing AI as a partner in the lexicographic workflow.

The papers collected in this workshop illustrate both dimensions of this transformation. Several contributions (Jung et al., Christofidou et al., Mathioudakis & Fiotaki, Ogrodniczuk & Czerski, Glebus, and Han & de Schryver) investigate how LLMs can support different stages of the lexicographic treatment of neologisms, ranging from candidate detection and corpus monitoring to definition writing, semantic analysis, validation, and editorial decision-making. Rather than viewing AI as a replacement for lexicographers, these studies consistently advocate not just human-in-the-loop but genuinely human-centric workflows in which corpus evidence, AI-generated analyses, and expert lexicographic judgement complement one another. The resulting picture is one of collaboration rather than automation, with lexicographers retaining

1 See https://globalex.link/workshops/ for an overview and links to individual workshops.
2 See https://eneoli.eu/

responsibility for interpreting linguistic evidence and ensuring the quality and reliability of dictionary content.

A second recurring theme concerns the evolution of lexicographic resources themselves. Several papers propose innovative models for representing lexical innovation, including discourse-oriented dictionary entries (Storjohann, and Cimander & Storjohann), graph-based lexical resources (Saputra & Barbosa), comparative frameworks for selecting candidate neologisms (Heylen et al.) with an awareness of potential AI-generated neologisms (Holmer), and a multilingual meta-terminological resource for neology studies and practice (Kernerman et al.). These contributions demonstrate that documenting lexical innovation requires more than adding new headwords to existing dictionaries. Instead, they argue for richer representations that capture semantic relations, discourse contexts, conceptual organisation, diachronic development, and a shared conceptual framework for the description of neology. At the same time, they show how AI technologies can facilitate the migration toward these more structured lexical knowledge bases by assisting in annotation, categorisation, relation extraction, and data enrichment.

This collection of abstracts also highlights that neology extends beyond general-language lexicography. Several contributions (Han & de Schryver, Ostroški et al., and Saputra & Barbosa) examine innovation in specialised languages and terminology, where conceptual change, multilingual communication, and standardisation raise additional methodological questions. The increasing interaction between lexicographic and terminological approaches illustrates that many of the challenges associated with AI, corpus-based detection and validation, and structured representation transcend traditional disciplinary boundaries.

Taken together, the contributions to GWLN-8 suggest that the role of lexicographers is not diminishing but becoming more extensive and complex. The lexicographer increasingly acts as a mediator between corpus-based empirical evidence, AI-generated analyses, evolving discourse communities, and structured lexical knowledge. Rather than replacing human expertise, AI amplifies the need for transparent editorial criteria, robust lexicographic methodologies, and interoperable lexical resources. These themes also resonate strongly with the objectives of ENEOLI and point towards a future in which lexicography, terminology, corpus linguistics and AI are increasingly integrated in the shared task of documenting lexical innovation.

We would like to thank all authors for their submissions, the members of the Scientific Committee for their careful reviews, the organisers of the XXII EURALEX International Congress for hosting GWLN-8, and the COST Action ENEOLI for its support. We hope that the papers will stimulate further discussion and collaboration on how lexicography can continue to research, interpret, and document lexical innovation in an era increasingly characterised by human–AI interaction. Selected papers will be published in the form of a thematic issue of the *International Journal of Lexicography*.

Contact information

Ilan Kernerman
K Dictionaries
ilan@lexicala.com

Kris Heylen
Dutch Language Institute
kris.heylen@ivdnt.org

Voula Giouli, Athina Sioupi, Nina Böbel & Alexander Ziem

CONSTRUCTICOGRAPHY IN PRACTICE

Constructions, Challenges, and Opportunities

Keywords Constructicography; construction grammar; computational lexicography; challenges and opportunities; NLP techniques; Constructicons

Constructicography (Lyngfelt et al., 2008), the systematic description of constructions in Constructicons and construction-based lexicographic resources, has emerged as a key area at the intersection of Construction Grammar (i.e., Croft, 2001), lexicography, and computational linguistics. While constructions have long been acknowledged as central units of linguistic knowledge, their consistent representation, annotation, and integration into lexicographic resources remain methodologically and technologically challenging.

The proposed workshop aims to bring together researchers working on theoretical, descriptive, and computational aspects of constructicography. It will focus on how constructions are identified, modeled, represented, and linked to lexical, semantic, and pragmatic information across languages and resource types. Particular attention will be paid to challenges such as construction granularity, variation, productivity, cross-linguistic comparability, corpus-driven discovery, and interoperability with existing lexical resources.

At the same time, the workshop highlights emerging opportunities offered by large corpora, annotation frameworks, and NLP techniques, including construction mining, semi-automatic Constructicon building, and the use of large language models in constructional analysis. Moreover, the workshop will explore applications of Constructicons, particularly in language teaching and learning.

By fostering dialogue between linguists, lexicographers, educational experts and computational researchers, the workshop seeks to advance constructicography as a mature and practically applicable field within lexicography.

References

Croft, W. (2001). *Radical Construction Grammar. Syntactic theory in typological perspective*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198299554.001.0001

Lyngfelt, B., Borin, L., Ohara, K. & Torrent, T.T. (Eds.) (2018). *Constructicography: Constructicon development across languages*. John Benjamins Publishing Company. https://doi.org/10.1075/cal.22

Contact information

Voula Giouli
Aristotle University of Thessaloniki
pgiouli@del.auth.gr

Athina Sioupi
Aristotle University of Thessaloniki
sioupi@del.auth.gr

Nina Böbel
Heinrich Heine University Düsseldorf
nina.boebel@hhu.de

Alexander Ziem
Heinrich Heine University Düsseldorf
ziem@hhu.de

Ondřej Matuška

POSTEDITING LEXICOGRAPHY WITH SKETCH ENGINE AND LEXONOMY

Keywords Post-editing workflows; Sketch Engine; lexonomy; corpus-based lexicography; AI-assisted dictionary making; Dictionary Express method

The tutorial aims to promote the understanding of state-of-the-art lexicographic practices based on the post-editing of corpus-generated and AI-generated content. It focuses on how contemporary lexicography can effectively combine automation with expert human judgement to achieve efficient, scalable, and cost-effective editorial workflows while ensuring that all dictionary data remain fully reviewed and validated by human editors.

The workshop provides an overview of current corpus and AI technologies relevant to dictionary making. Sketch Engine (https://www.sketchengine.eu/) is exploited to support key lexicographic tasks such as sense analysis, collocation extraction, example selection, and pattern discovery. The session will also introduce Lexonomy (https://guide.lexonomy.eu/) as a modern dictionary writing system designed for updating legacy dictionaries as well as creating new lexicographic works. Lexonomy demonstrates how corpus evidence and post-edited content can be efficiently structured, edited, and published within a collaborative editorial environment.

A central component of the tutorial is the presentation of the Dictionary Express method (https://dictionary.express/), which combines automated content generation with systematic post-editing to **accelerate dictionary production** without compromising quality. Through practical workflows and concrete examples, the tutorial will show how corpus data, AI-assisted suggestions, and editorial guidelines can be integrated into a coherent process that supports consistency, transparency, and editorial control. The session is intended for lexicographers, editors, and researchers interested in applying corpus-based and AI-supported methods to contemporary dictionary projects.

Contact information

Ondřej Matuška
Lexical Computing
ondrej.matuska@sketchengine.eu

Lian Chen & Besim Kabashi

MULTI-WORD EXPRESSIONS AND PHRASEOLOGY

Corpus-Based and Computer-Processed

Keywords Phraseology; multi-word expressions; computational lexicography; corpus linguistics; lexicographic resources; AI-assisted lexicography

Phraseology (Cowie, 1998; Granger & Meunier, 2008; Mitkov, 2017; Mel'čuk, 2023; Polguère, 2002, 2014; Mejri, 2018; Chen, 2021), or multiword expressions (MWEs) (Savary, 2008; Constant, 2012) occupy a central position in linguistic description, language use, and language learning. Idioms, collocations, lexical bundles, and fixed or semi-fixed expressions constitute a significant part of natural language, yet they remain challenging to model, describe, and represent in lexicographic resources (Pecina, 2010; Mel'čuk, 2011; Polguère, 2014; Chen, 2025). With the rapid development of new technologies—particularly corpus linguistics, Natural Language Processing (NLP), and Artificial Intelligence (AI)—phraseology and lexicography are currently undergoing a profound methodological and conceptual transformation.

From a modeling perspective, new technologies make it possible to move beyond traditional, intuition-based descriptions of phraseological units. Large-scale corpora (Mitkov, 2017), both general and specialized, enable the systematic identification of MWEs through statistical, distributional, and syntactic approaches. Methods such as n-gram extraction, association measures (PMI, t-score), syntactic patterning, and embedding-based similarity allow researchers to capture degrees of fixedness, semantic compositionality, and contextual variability. These advances open new perspectives for representing phraseological knowledge in structured models, including ontologies, lexical networks, and standards such as OntoLex-Lemon (McCrae et al., 2017; Bosque-Gil et al., 2019).

In terms of resource creation, technological tools have profoundly reshaped lexicographic practices (Atkins & Rundell, 2008; Granger & Paquot, 2012). Digital corpora, web-based data, and annotation platforms facilitate the semi-automatic extraction and validation of phraseological units (Mitkov, 2017; Evert, 2008; Gries, 2008; Constant et al., 2017). New-generation lexical resources integrate rich metadata, usage examples, frequency information, and semantic relations, making them more dynamic and interoperable (Polguère, 2014; Cimiano et al., 2016; McCrae et al., 2017). Moreover, computational approaches enable the development of multilingual and contrastive resources, which are essential for studying phraseological variation across languages and cultures (Paquot, 2015; Mel'čuk, 2011).

New technologies also play a crucial role in the design of pedagogical dictionaries and learning-oriented resources (Bogaards & van der Kloot, 2001; Lew, 2012). Phraseology is often a major obstacle for language learners, as MWEs cannot always be interpreted compositionally (Wray, 2002; Howarth, 1998; Granger, 1998). Corpus-based examples, learner-oriented definitions, and adaptive digital interfaces can significantly enhance the accessibility and usability of phraseological information (Carter & Sinclair, 2004; Paquot, 2015). AI-driven tools, such as intelligent tutoring systems or LLM-assisted lexicography, offer promising avenues for generating contextualized examples, usage notes, and difficulty pro-files tailored to learners' needs (Heift & Schulze, 2007; Godwin-Jones, 2023; Bender & Koller, 2020).

In translation lexicography, the impact of new technologies is equally significant (Hartmann, 2007; Atkins & Rundell, 2008). Phraseological units pose well-known challenges in translation

due to their idiomaticity and phraseocultural specificity (Chen, 2022a, 2022b). Parallel corpora, alignment tools, and machine translation systems provide valuable data for identifying translation equivalents, partial correspondences, and translation strategies. Digital translation dictionaries can now incorporate cross-linguistic mappings, semantic annotations, and real attested examples, bridging the gap between lexicographic description and actual translational practice.

Finally, the analysis of existing dictionaries through technological means (Béjoint, 2010) offers new insights into lexicographic traditions and practices. Computational methods allow for the large-scale comparison of dictionary entries, coverage, microstructure, and treatment of phraseological units. Such analyses contribute to a critical understanding of how dictionaries evolve in response to technological, linguistic, and societal changes (Hausmann, 1989; Fuertes-Olivera & Bergenholtz, 2011).

The intersection of phraseology, multi-word expressions, lexicography, and new technologies constitutes a fertile research domain, fostering innovative models, richer resources, and more effective tools for analysis, learning, and translation.

References

Atkins, B. T. S. & Rundell, M. (2008). *The Oxford guide to practical lexicography*. Oxford University Press.

Bender, E. M. & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In D. Jurafsky, J. Chai, N. Schluter & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5185–5198). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.463

Béjoint, H. (2010). *The lexicography of English: From origins to present.* Oxford University Press.

Bogaards, P. & van der Kloot, W. A. (2001). The use of grammatical information in learners' dictionaries. *International Journal of Lexicography*, 14(2), 97–121. https://doi.org/10.1093/ijl/14.2.97

Bosque-Gil, J., & Gracia, J. (Eds.) (2019). *The OntoLex Lemon Lexicography Module*. Final Community Group Report. World Wide Web Consortium (W3C).

Carter, R. & Sinclair, J. (2004). *Trust the text: Language, corpus and discourse*. Routledge.

Cimiano, P., McCrae, J. P., & Buitelaar, P. (2016). *Lexicon Model for Ontologies*: Community Report.

Chen, L. (2021). *Analyse comparative des expressions idiomatiques en chinois et en français* [Doctoral dissertation, CY Cergy Paris Université].

Chen, L. (2022a). Phraseoculture in the construction of the corpus of the DiCoP: The treatment of the phraseographic microstructure. In *Proceedings of the International Conference EUROPHRAS 2022* (pp. 17–25). https://doi.org/10.26615/978-954-452-080-9_003

Chen, L. (2022b). Phraséoculturologie: Une sous-discipline moderne indispensable de la phraséologie. *SHS Web of Conferences*, 138, 04011. https://doi.org/10.1051/shsconf/202213804011

Chen, L. (2025). Modeling and structuring of a bilingual French–Chinese phraseological dictionary: Neural automatic approach for ontology and lexicography. In Kosem, I. (Ed.), *eLex 202: Electronic lexicography in the 21st century Intelligent Lexicography* (pp. 830–851). Bled, Slovenia.

Constant, M., Eryiğit, G., Monti, J., van der Plas, L., Ramisch, C., Rosner, M. & Todirascu, A. (2017). Multiword expression processing: A Survey. *Computational Linguistics*, 43(4), 837–892.

Constant, M. (2012). Mettre les expressions multi-mots au cœur de l'analyse automatique de textes: sur l'exploitation de ressources symboliques externes. Traitement du texte et du document. Université Paris-Est. ⟨tel-00841556⟩.

Cowie, A. P. (1998). *Phraseology: Theory, analysis, and applications*. Oxford University Press.

Evert, S. (2008). Corpora and collocations. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics*. An international Handbook (pp. 1212–1248). Mouton de Gruyter.

Fuertes-Olivera, P. A., & Bergenholtz, H. (2011). *e-Lexicography: The Internet, Digital Initiatives and Lexicography*. Bloomsbury.

Godwin-Jones, R. (2023). Generative AI and language learning. *Language Learning & Technology*, 27(2), 4–31.

Granger, S. (1998). Prefabricated patterns in advanced EFL writing: Collocations and formulae. In A. P. Cowie (Ed.), *Phraseology: Theory, analysis, and applications* (pp. 145–160). Oxford University Press.

Granger, S., & Meunier, F. (Eds.) (2008). *Phraseology: An interdisciplinary perspective*. John Benjamins. https://doi.org/10.1093/oso/9780198294252.003.007

Granger, S. & Paquot, M. (Eds.) (2012). *Electronic Lexicography*. Oxford University Press.

Gries, S. T. (2008). Phraseology and linguistic theory. In S. Granger & F. Meunier (Eds.), *Phraseology: An interdisciplinary perspective* (pp. 3–26). John Benjamins.

Hartmann, R. R. K. (2007). *Interlingual lexicography: Selected Essays on Translation Equivalence, Contrastive Linguistics, and the Bilingual Dictionary*. Niemeyer.

Hausmann, F. J. (1989). Die gesellschaftlichen Aufgaben der Lexikographie. In F. J. Hausmann, O. Reichmann, H. E. Wiegand & L. Zgusta (Eds.), *Wörterbücher* (pp. 1–19). De Gruyter.

Heift, T., & Schulze, M. (2007). *Errors and intelligence in Computer-Assisted Language Learning: Parsers and Pedagogues*. Routledge.

Howarth, P. (1998). Phraseology and second language proficiency. *Applied Linguistics*, 19(1), 24–44.

Lew, R. (2012). How can we make electronic dictionaries more effective? In S. Granger & M. Paquot (Eds.), *Electronic lexicography* (pp. 343–362). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199654864.003.0016

McCrae, J. P., 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: Proceedings of eLex* 2017 (pp. 587–597). Lexical Computing CZ.

Mel'čuk, I. A. (2011). Phrasèmes dans le dictionnaire. In J.-C. Anscombre & S. Mejri (Eds.), *Le figement linguistique: La parole entravée* (pp. 41–62). Honoré Champion.

Mel'čuk, I. A. (2023). *General phraseology: Theory and practice*. John Benjamins.

Mejri, S. (2018). La phraséologie: Cotexte, contexte et contenus culturels. *Lublin Studies in Modern Languages and Literature*, 42(4), 11–38. https://doi.org/10.17951/lsmll.2018.42.4.11

Mitkov, R. (Ed.) (2017). *Computational and corpus-based phraseology. Second International Conference, Europhras 2017, London, UK, November 13-14, 2017, Proceedings*. Springer.

Pecina, P. (2010). Lexical association measures and collocation extraction. *Language Resources and Evaluation*, 44(1), 137–158. https://doi.org/10.1007/s10579-009-9101-4

Paquot, M. (2015). Lexicography and phraseology. In Biber, D. & R. Reppen (Eds.), *The Cambridge Handbook of Corpus Linguistics* (pp. 460-477). Cambridge University Press.

Polguère, A. (2002). *Notions de base en lexicologie*. Ophrys.

Polguère, A. (2014). Principes de modélisation systémique des réseaux lexicaux. *Proceedings of TALN 2014* (pp. 79–90). Association pour le Traitement Automatique des Langues.

Savary, A. (2008). Computational inflection of multi-word units: A contrastive study of lexical approaches. *Linguistic Issues in Language Technology*, 1, 1–53. https://doi.org/10.33011/lilt.v1i.1195

Wray, A. (2002). *Formulaic Language and the Lexicon*. Cambridge University Press.

Contact information

Lian Chen
Laboratoire LLL, University of Orléans & CRLAO–CNRS–INALCO, France
loselychen@gmail.com

Besim Kabashi
Eberhard Karls Universität Tübingen, Germany
besim.kabashi@fau.de

INDEX OF AUTHORS

EURALEX 2026

LIST OF AUTHORS

Author name	Email	Country	Affiliation
Esra Abdelzaher	esra.abdelzaher@gmail.com	Hungary	Department of English Linguistics, University of Debrecen
Andrea Abel	andrea.abel@eurac.edu	Italy	Eurac Research
Vassiliki Afentoulidou	vafentoul@phil.uoa.gr	Greece	National and Kapodistrian University of Athens & Academy of Athens
Luke Omoyemi Akinremi	lukeakinremi@gmail.com	Spain	Universidade de Santiago de Compostela
Abolfazl Alamdar	abalfazl.alamdar@yahoo.com	Iran	PhD, Linguistics, Allameh Tabataba'i University, Tehran, Iran
Mikel Alonso	mikelalon@gmail.com	Spain	University of the Basque Country
Maria Ammari	maria.ammari@amu.edu.pl	Poland	Adam Mickiewicz University
Jinsan An	san@knu.ac.kr	South Korea	Kyungpook National University
Tjaša Arčon	tjasa.arcon@fr.uni-lj.si	Slovenia	University of Ljubljana
Iván Arias-Arias	ivanarias.arias@usc.gal	Spain	Universidade de Santiago de Compostela
Blanca Arias-Badia	blanca.arias@upf.edu	Spain	Pompeu Fabra University
Petra Bago	pbago@m.ffzg.hr	Croatia	University of Zagreb
Martina Bajčić	martina.bajcic@pravri.uniri.hr	Croatia	University of Rijeka
Michela Bandini	michela.bandini@ilc.cnr.it	Italy	Institute Of Computational Linguistics “A. Zampolli”
Sílvia Barbosa	silviabarbosa@fcsh.unl.pt	Portugal	NOVA FCSH CLUNL
Andreas Baumann	andreas.baumann@univie.ac.at	Austria	University of Vienna
Slobodan Beliga	sbeliga@uniri.hr	Croatia	University of Rijeka
Andrea Bellandi	andrea.bellandi@ilc.cnr.it	Italy	Institute Of Computational Linguistics “A. Zampolli”
Elisenda Bernal	elisenda.bernal@upf.edu	Spain	Pompeu Fabra University
Elena Berthemet	elenaberthemet@free.fr	France	Centre de Linguistique en Sorbonne
Harald Bichlmeier	harald.bichlmeier@uni-jena.de	Germany	Sächsische Akademie der Wissenschaften zu Leipzig

Author name	Email	Country	Affiliation
Riccardo Billero	rbillero@unibz.it	Italy	Free University of Bozen-Bolzano
Black Bear Jr., Ben	benblkbr@lakhota.org	United States	Lakota Language Consortium
Marek Blahuš	marek.blahus@sketchengine.eu, marek.blahus@sketchengine.co.uk	Czechia	Lexical Computing
Nina Böbel	nina.boebel@hhu.de	Germany	Heinrich Heine University Düsseldorf
Igor Boguslavsky	igor.m.boguslavsky@gmail.com	Spain	Universidad Politécnica de Madrid
Onorina Botezat	onorina.botezat@rei.ase.ro	Romania	Bucharest University of Economic Studies
Theo J.D. Bothma	theo.bothma@up.ac.za	South Africa	University of Pretoria
Biljana Božinovski	biljana.bozinovski@um.si	Slovenia	University of Maribor
Ivana Brač	ibrac@ihjj.hr	Croatia	Institute of Croatian Language
Katja Brankačkec	katja.brankackec@slu.cas.cz	Czechia	Slovanský Ústav Akademie věd ČR
Cormac Breathnach	cbreathnach@forasnagaeilge.ie	Ireland	Foras na Gaeilge
Marta Brescia-Zapata	marta.brescia@upf.edu	Spain	Pompeu Fabra University
Gabriela Brůhová	gabriela.bruhova@ff.cuni.cz	Czechia	Charles University
Rukiye Smilla Burkart	burkart@uni-trier.de	Germany	Trier University
John J. Camilleri	john.j.camilleri@chalmers.se	Sweden	Chalmers University of Technology / University of Gothenburg
Edmond Cane	ecane2000@yahoo.com	Albania	Luarasi University
Valeria Caruso	vcaruso@unior.it	Italy	Università degli Studi di Napoli L'Orientale
Elena Castaldi	castaldi.elena2003@gmail.com	Italy	EMLex
Elina Chadjipapa	elinaxp@hotmail.com	Greece	Democritus University of Thrace
Lian Chen	loselychen@gmail.com	France	LLL, University of Orleans
John J. Chew III	poslfit@gmail.com	Canada	Canadian English Dictionary
Dalí Chirino	dali.chirino@ivdnt.org	Netherlands	Instituut voor de Nederlandse Taal
Anastasia Christofidou	christo@academyofathens.gr	Greece	National and Kapodistrian University of Athens & Academy of Athens

Author name	Email	Country	Affiliation
Luisa Cimander	cimander@ids-mannheim.de	Germany	Leibniz-Institut für Deutsche Sprache, Mannheim
Lut Colman	lut.collman@ivdnt.org	Netherlands	Instituut voor de Nederlandse Taal
Li Cui	choiri20@yonsei.ac.kr	South Korea	Yonsei University
Lidija Cvikić	lidija.cvikic@ufzg.hr	Croatia	University of Zagreb
Václav Cvrček	vaclav.cvrcek@ff.cuni.cz	Czechia	Charles University
Dariusz Czerski	dariusz.czerski@ipipan.waw.pl	Poland	Institute of Computer Science, Polish Academy of Sciences
Tomasz Czyż	tc86861@student.uwb.edu.pl	Poland	University of Białystok
Huy-Linh Dao	huy-linh.dao@inalco.fr	France	CNRS-CRLAO
Jesse de Does	Jesse.Dedoes@ivdnt.org	Netherlands	Dutch Language Institute
Anna De Meo	ademeo@unior.it	Italy	Università degli Studi di Napoli L'Orientale
Flávia de Oliveira Maia-Pires	projetosmaiapires@gmail.com	Brazil	University of Brasília (UnB)
Gilles-Maurice de Schryver	gillesmaurice.deschryver@UGent.be	Belgium	Ghent University
Veronique De Tier	veronique.detier@ivdnt.org	Belgium	Dutch Language Institute
Janina Deilke	Janina.Deilke@uibk.ac.at	Austria	University of Innsbruck
Clara Anush Del Rey Adanalian	claraanush.delrey@upf.edu	Spain	Pompeu Fabra University
Alexander Delaporte	alexander.delaporte@ehess.fr	France	CNRS-CRLAO
Katrien Depuydt	katrien.depuydt@ivdnt.org	Netherlands	Dutch Language Institute
Kristina Š. Despot	kdespot@ihjj.hr	Croatia	Institute of Croatian Language
Vyacheslav Dikonov	dikonov@iitp.ru	Russia	Russian Academy of Science
Milica Dinić Marinković	milica.dinic.marinkovic@fil.bg.ac.rs	Serbia	University of Belgrade
María José Domínguez Vázquez	majo.dominguez@usc.es	Spain	Universidade de Santiago de Compostela
Alona Dorozhynska	alonochkatkachuk@gmail.com	Ukraine	Ukrainian Lingua-Information Foundation of NAS of Ukraine
Ifigeneia Dosi	dosi.if@gmail.com	Greece	Democritus University of Thrace
Ida Dringo-Horvath	dringo.horvath.ida@kre.hu	Hungary	Karoli Gaspar University of the Reformed Church in Hungary
Luca Ducceschi	luca.ducceschi@eurac.edu	Italy	Eurac Research

Author name	Email	Country	Affiliation
Elisabeth Eder	elisabeth.eder@oeaw.ac.at	Austria	Austrian Academy of Sciences
Dwayne Ellul	dwayne.ellul@um.edu.mt	Malta	L-Università ta' Malta
Daniel Elsner	Daniel.Elsner@oeaw.ac.at	Austria	Austrian Academy of Sciences
Ivana Eterović	isankovi@m.ffzg.hr	Croatia	University of Zagreb
Mathieu Fannee	mathieu.fannee@ivdnt.org	Netherlands	Dutch Language Institute
Emma Ferrett	emma.ferrett@queensu.ca	Canada	Queen's University
Sophia Fiedler	sophia.fiedler@ids-mannheim.de	Germany	Leibniz-Institut für Deutsche Sprache
Ivana Filipović Petrović	ivana.filipov@gmail.com	Croatia	Croatian Academy of Sciences and Arts
Irene Fioravanti	irene.fioravanti@unistrapg.it	Italy	University for Foreigners of Perugia
Alexandra Fiotaki	alexandra.fiotaki@gmail.com	Greece	University of Athens
Hanna Fischer	hanna.fischer@uni-marburg.de	Germany	Research Center Deutscher Sprachatlas
Ida Flörke	if@dsl.dk	Denmark	Det Danske Sprog- og Litteraturselskab
Emmanuel Sylvain Fomat	sylvainfomat4@gmail.com, fomat@uni-hildesheim.de, emmanuelsylvain.format@rai.usc.es	Germany	University of Hildesheim
Daniele Franceschi	daniele.franceschi72@gmail.com	Italy	Roma Tre University
Mihaela Franjić	mihaela.franjic1@um.si	Slovenia	University of Maribor
Ana Frankenberg-Garcia	a.frankenberg-garcia@surrey.ac.uk	United Kingdom	University of Surrey
Judit Freixa	judit.freixa@upf.edu	Spain	Universitat Pompeu Fabra
Margarita Freixas	Margarita.Freixas@uab.cat	Spain	Universitat Autònoma de Barcelona
Tatiana Frolova	tfrolova@gmail.com	Russia	Russian Academy of Science
M.ª Teresa Fuentes Morán	tfuentes@usal.es	Spain	Universidad de Salamanca
Dejan Gabrovšek	dejan.gabrovsek@zrc-sazu.si	Slovenia	ZRC SAZU, Fran Ramovš Institute of the Slovenian Language
Christine Ganslmayer	Christine.Ganslmayer@fau.de	Germany	University of Erlangen–Nuremberg
Polona Gantar	apolonija.gantar@guest.arnes.si	Slovenia	University of Ljubljana
Marta Garcia Casado	marta.garciac@upf.edu	Spain	Pompeu Fabra University

Author name	Email	Country	Affiliation
Pol Garriga Martínez	pol.garriga@upf.edu	Spain	Pompeu Fabra University
Anne-Kathrin Gärtig-Bressan	akgaertig@units.it	Italy	University of Trieste
Zoe Gavriilidou	zoegab@otenet.gr, zoegabriil@gmail.com	Greece	Democritus University of Thrace
Alexander Geyken	geyken@bbaw.de	Germany	Berlin-Brandenburg Academy of Sciences and the Humanities
Amir Ghorvei	17mag@queensu.ca	Canada	Queen's University
Aretousa Giannakou	giannakou.aretousa@ac.eap.gr	Greece	Academy of Athens
Voula Giouli	pgiouli@del.auth.gr	Greece	Aristotle University of Thessaloniki
Emiliano Giovannetti	emiliano.giovannetti@ilc.cnr.it	Italy	Institute Of Computational Linguistics “A. Zampolli”
Marius Glebus	marius.glebus@lki.lt	Lithuania	Institute of the Lithuanian Language
Peter Gobel	pgobel@cc.kyoto-su.ac.jp	Japan	Kyoto Sangyo University
Rufus H. Gouws	rhg@sun.ac.za	South Africa	Stellenbosch University
Rita Gramellini	rita.gramellini2@unibo.it	Italy	Alma Mater Studiorum Università di Bologna
Mikus Grasmanis	mikus.grasmanis@lumii.lv	Latvia	University of Latvia
Jan Hajič	hajic@ufal.mff.cuni.cz	Czechia	Charles University
Wanjing Han	Wanjing.Han@UGent.be	Belgium	Ghent University
Ulrich Heid	heidul@uni-hildesheim.de	Germany	University of Hildesheim
Barbara Heinisch	barbara.heinisch@univie.ac.at	Italy	Eurac Research
Tarja Heinonen	tarja.heinonen@kotus.fi	Finland	Institute for the Languages of Finland (Kotimaistenkielten keskus)
Ondřej Herman	ondrej.herman@sketchengine.co.uk	Czechia	Lexical Computing
Esther Hernández	esther.hernandez@cchs.csic.es	Spain	Agencia Estatal Consejo Superior de InvestigacionesCientíficas (CSIC)
Kris Heylen	Kris.Heylen@kuleuven.be	Netherlands	Dutch Language Institute
Louise Holmer	louise.holmer@svenska.gu.se	Sweden	Språkbanken, Department of Swedish, Multilingualism, Language Technology University of Gothenburg

Author name	Email	Country	Affiliation
Marijana Horvat	mhorvat@ihjj.hr	Croatia	Institute of Croatian Language
Martina Horvat	mshorvat@ihjj.hr	Croatia	Institute of Croatian Language
Lana Hudeček	lhudecek@ihjj.hr	Croatia	Institute of Croatian Language
John Humbley	humbley.john@orange.fr	France	Université Paris Cité
Polina Iaroshenko	polina.iaroshenko@yandex.ru	Russia	Lomonosov Moscow State University
Paula Igareda González	paula.igareda@upf.edu	Spain	Pompeu Fabra University
Evgeniya Inshakova	e.s.inshakova@gmail.com	Russia	Russian Academy of Science
Miloš Jakubíček	milos.jakubicek@sketchengine.eu	Czechia	Lexical Computing
Laura A. Janda	laura.janda@uit.no	Norway	The Arctic University of Norway
Janoš Ježovnik	janos.jezovnik@zrc-sazu.si	Slovenia	ZRC SAZU, Fran Ramovš Institute of the Slovenian Language
Daniel Jiménez Casas	daniel.jimenezcasas@upf.edu	Spain	Pompeu Fabra University
Ellert Thor Johannsson	etj@hi.is	Iceland	The Árni Magnússon Institute for Icelandic Studies
Željko Jozić	zjozic@ihjj.hr	Croatia	Institute of Croatian Language
Hae-Yun Jung	haeyun.jung.22@gmail.com	South Korea	Kyungpook National University
Besim Kabashi	besim.kabashi@fau.de	Germany	Eberhard Karls University of Tübingen
Susanne Kabatnik	kabatnik@uni-trier.de	Germany	Trier University
Jelena Kallas	jelena.kallas@eki.ee	Estonia	Institute of the Estonian Language
Makimi Kano	makimik@cc.kyoto-su.ac.jp	Japan	Kyoto Sangyo University
Boris Kern	boris.kern@zrc-sazu.si	Slovenia	Research Centre of the Slovenian Academy of Sciences and Arts / University of Nova Gorica
Ilan Kernerman	ilan@lexicala.com	Israel	K Dictionaries
Václava Kettnerová	kettnerova@ufal.mff.cuni.cz	Czechia	Charles University
Ahto Kiil	ahto.kiil@eki.ee	Estonia	Institute of the Estonian Language
Agnes Kim	agnes.kim@osu.cz	Czechia	University of Ostrava
Matej Klemen	matej.klemen@fri.uni-lj.si	Slovenia	University of Ljubljana
Veronika Kolářová	kolarova@ufal.mff.cuni.cz	Czechia	Charles University
Evanthia Konstantinidou	evi1990@hotmail.gr	Greece	Democritus University of Thrace

Author name	Email	Country	Affiliation
Kristina Koppel	kristina.koppel@eki.ee	Estonia	Institute of the Estonian Language
Wolfgang Koppensteiner	Wolfgang. Koppensteiner@oeaw.ac.at	Austria	Austrian Academy of Sciences
Mariia Kopylova	mashulya.kopylova@list.ru	Italy	Roma Tre University
Katharina Korecky-Kröll	katharina.korecky-kroell@ oeaw.ac.at	Austria	Austrian Academy of Sciences
Iztok Kosem	iztok.kosem@cjvt.si	Slovenia	University of Ljubljana
Vojtěch Kovář	vojtech.kovar@sketchengine.eu	Czechia	Lexical Computing
František Kovařík	frantisek.kovarik@ sketchengine.eu	Czechia	Masaryk University
Dominika Kováříková	dominika.kovarikova@ff.cuni.cz	Czechia	Charles University
Iwona Krawczyk	iwona.krawczyk@ijp.pan.pl	Poland	Polish Academy of Sciences
Simon Krek	simon.krek@guest.arnes.si	Slovenia	Jožef Stefan Institute
Martin Kroon	m.s.kroon@hum.leidenuniv.nl	Netherlands	Instituut voor de Nederlandse Taal
Cvetana Krstev	CvetanaJK@gmail.com	Serbia	JeRTeh Society
Theresa Kruse-Kurbach	kruset@uni-hildesheim.de	Germany	University of Hildesheim
Domen Krvina	domen.krvina@zrc-sazu.si	Slovenia	ZRC SAZU, Fran Ramovš Institute of the Slovenian Language
Katarzyna Kryńska	katarzyna.erhardt@gmail.com	Poland	Polish Academy of Sciences
Amila Kugic	amila.kugic@medunigraz.at	Austria	Medical University of Graz
Tamar Laluashvili	tamar.laluashvili.1@iliauni.edu.ge	Georgia	Ilia State University
Alfred Lameli	lameli@uni-marburg.de	Germany	Research Center Deutscher Sprachatlas
Margit Langemets	margit.langemets@eki.ee	Estonia	Institute of the Estonian Language
Alexandre Lazursky	lazursky@mail.ru	Russia	Russian Academy of Science
Nina Ledinek	ledinekn@gmail.com	Slovenia	ZRC SAZU, Fran Ramovš Institute of the Slovenian Language
Seulki Lee	nono0118@naver.com	South Korea	Chungbuk National University
Jun Lee	leejun0624@gmail.com	South Korea	Yonsei University
Robert Lew	rlew@amu.edu.pl	Poland	Adam Mickiewicz University
Barbara Lewandowska-Tomaszczyk	barbara.lewandowska-tomaszczyk@konin.edu.pl	Poland	State University of Applied Sciences in Konin

Author name	Email	Country	Affiliation
Silvio Licurgo Pires	licurgoslp@gmail.com	Brazil	University Center of Brasília (UniCEUB)
David Lindemann	david.lindemann@ehu.eus	Spain	EHU University of the Basque Country
Pia Carmela Lombardi	piacarmela.lombardi@units.it	Italy	Università degli Studi di Trieste
Henrik Lorentzen	hl@dsl.dk	Denmark	Society for Danish Language and Literature
Natalia Loukachevitch	louk_nat@mail.ru	Russia	Research Computing Center of Moscow State University
Ivana Lovrić Jović	ilovric@ihjj.hr	Croatia	Institute of Croatian Language
Liina Lutsepp	llutsepp@tlu.ee	Estonia	Tallinn University
Tibor M. Pintér	m.pinter.tibor@kre.hu	Hungary	Károli Gáspár University of the Reformed Church in Hungary
Hanna Maask	hanna.maask@eki.ee	Estonia	Institute of the Estonian Language
Magdalena Majdak	magdalena.majdak@ijp.pan.pl	Poland	Polish Academy of Sciences
Marine Makhatadze	marine_makhatadze@yahoo.com	Georgia	Ivane Javakhishvili Tbilisi State University
Tinatin Margalitadze	tinatin.margalitadze@iliauni.edu.ge	Georgia	Ilia State University
Saša Marjanović	sasa.marjanovic@fil.bg.ac.rs	Serbia	University of Belgrade
Anna Marklová	a.marklova@gmail.com	Czechia	Charles University
Aleksandra Marković	aleksandra.markovic@isj.sanu.ac.rs	Serbia	Serbian Academy of Sciences and Arts
Jagoda Marszałek	jagoda.marszalek@ijp.pan.pl, yagoodka2@gmail.com	Poland	Polish Academy of Sciences
Elena Martín-Cancela	elena.martin@udc.es	Spain	Universidade da Coruña
Leonor Martins	leonor.olivmartins22@gmail.com	Portugal	Academy of Sciences of Lisbon
Abdelouahab Masbah	masbah.abdelouahab@gmail.com	Italy	Institute Of Computational Linguistics “A. Zampolli”
Nikos Mathioudakis	nikosmathious@gmail.com	Greece	Research Centre for Scientific Terms and Neologisms, Academy of Athens
Ondřej Matuška	ondrej.matuska@sketchengine.eu	Czechia	Lexical Computing
Ketevan Mchedlishvili	ketevani.mchedlishvili.3@iliauni.edu.ge	Georgia	Ilia State University
Nathalie Mederake	mederake@staff.uni-marburg.de	Germany	Research Center Deutscher Sprachatlas

Author name	Email	Country	Affiliation
Cordula Meißner	cordula.meissner@uibk.ac.at	Austria	University of Innsbruck
Giorgi Meladze	giorgi.meladze.4@iliauni.edu.ge	Georgia	Ilia State University
Olga Menagarishvili	olga.menagarishvili@ce.gatech.edu	United States	Georgia Institute of Technology
Koen Mertens	koen.mertens@ivdnt.org	New Zealand	Dutch Language Institute
Peter Meyer	meyer@ids-mannheim.de	Germany	Leibniz-Institut für Deutsche Sprache (IDS)
Mija Michelizza	mija.michelizza@zrc-sazu.si	Slovenia	ZRC SAZU, Fran Ramovš Institute of the Slovenian Language
Tomasz Michta	t.michta@gmail.com	Poland	University of Bialystok
Gregor Middell	gregor.middell@bbaw.de	Germany	Berlin-Brandenburg Academy of Sciences and the Humanities
Ana Mihaljević	amihaljevi@gmail.com	Croatia	Old Church Slavonic Institute, Zagreb
Josip Mihaljević	jmihaljevic@stin.hr	Croatia	Old Church Slavonic Institute, Zagreb
Dorota Mika	dorota.mika@ijppan.pl	Poland	Polish Academy of Sciences
Agnes Aggergaard Mikkelsen	aam@dsl.dk	Denmark	Det Danske Sprog- og Litteraturselskab
Alba Milà Garcia	alba.mila@upf.edu	Spain	Pompeu Fabra University
Dejana Mirković-Birtašić	dejana.mirkovic.birtasic@fil.bg.ac.rs	Serbia	University of Belgrade
Jiří Mírovský	mirovsky@ufal.mff.cuni.cz	Czechia	Charles University
Maria Mitsiaki	mitsiakim@gmail.com	Greece	Democritus University of Thrace
Julie Moore	julie@juleswords.co.uk	United Kingdom	A.S. Hornby Trust
Diana Movsisyan	movsisyan.diana@gmail.com	Armenia	Armenian State University of Economics
Katarzyna Mroczyńska	katarzyna.mroczynska@gmail.com	Poland	University of Siedlce
Bruno Nahod	bnahod@ihjj.hr	Croatia	Institute of Croatian Language
Kilim Nam	nki@yonsei.ac.kr	South Korea	Yonsei University
Seol Namkung	sul2001sul@naver.com	South Korea	Yonsei University
Gunta Nešpore-Bērzkalne	gunta.nespore@gmail.com	Latvia	University of Latvia
Phoebe Nicholson	phoebe.nicholson3@oup.com	United Kingdom	Oxford University Press

Author name	Email	Country	Affiliation
Martina Nied Curcio	martina.nied@uniroma3.it	Italy	Università degli Studi Roma Tre
Sanni Nimb	sn@dsl.dk	Denmark	Det Danske Sprog- og Litteraturselskab (DSL)
Nathalie Norman	naha@hum.ku.dk	Denmark	University of Copenhagen
Damien Nouvel	damien.nouvel@inalco.fr	France	ERTILM
Krzysztof Nowak	krzysztof.nowak@ijppan.pl	Poland	Polish Academy of Sciences
Pádraig Ó Mianáin	pomianain@forasnagaeilge.ie	Ireland	Foras na Gaeilge
Maciej Ogrodniczuk	maciej.ogrodniczuk@ipipan.waw.pl	Poland	Institute of Computer Science, Polish Academy of Sciences
Giorgi Okropiridze	giorgi.okropiridze.4@iliauni.edu.ge	Georgia	Ilia State University
Emma Josefin Ölander Aadland	emma.aadland@uib.no	Norway	University of Bergen Library
Michal Olbrich	olbrich@ufal.mff.cuni.cz	Czechia	Charles University
Iryna Ostapova	irinaostapova@gmail.com	Ukraine	Ukrainian Lingua-Information Foundation of NAS of Ukraine
Ana Ostroški Anić	aostrosk@ihjj.hr	Croatia	Institute of Croatian Language
Katalin P. Márkus	p.markus.kata@kre.hu	Hungary	Karoli Gaspar University of the Reformed Church in Hungary
Christos Papanagiotou	cpapanagiotou@academyofathens.gr	Greece	Academy of Athens
Geda Paulsen	geda.paulsen@eki.ee	Estonia	Institute of the Estonian Language
Martina Pavić	mpavic@ihjj.hr	Croatia	Institute of Croatian Language
Benedikt Perak	bperak@uniri.hr	Croatia	University of Rijeka
Andrej Perdih	andrej.perdih@zrc-sazu.si	Slovenia	Research Centre of the Slovenian Academy of Sciences and Arts
Katarína Petreková	katarina.petrekova@sketchengine.eu	Czechia	Lexical Computing
Špela Petric Žižić	spela.petric@zrc-sazu.si	Slovenia	ZRC SAZU, Fran Ramovš Institute of the Slovenian Language
Silvia Piccini	silvia.piccini@ilc.cnr.it	Italy	Institute Of Computational Linguistics “A. Zampolli”
Klára Pivoňková	klara.pivonkova@ff.cuni.cz	Czechia	Charles University

Author name	Email	Country	Affiliation
Dóra Pődör	podor.dora@kre.hu	Hungary	Károli Gáspár University of the Reformed Church in Hungary
Esta Prangel	esta.prangel@eki.ee	Estonia	Institute of the Estonian Language
Linda Prossliner	Linda.Prossliner@uibk.ac.at	Austria	University of Innsbruck
Manuel Raaf	raaf@badw.de	Germany	Bavarian Academy of Sciences and Humanities
Anna-Lena Randermann	Anna-Lena.Randermann@uibk.ac.at	Austria	University of Innsbruck
Veronika Raušová	veronika.rausova@ff.cuni.cz	Czechia	Charles University
Laura Rebosio	Laura.Rebosio@uibk.ac.at	Austria	Universität Innsbruck
Geraint Paul Rees	geraintpaul.rees@upf.edu	Spain	Pompeu Fabra University
Anastasia Riehl	riehla@queensu.ca	Canada	Queen's University
Lydia Risberg	lydia.risberg@gmail.com	Estonia	Estonian Language Institute / University of Tartu
Ewa Rodek	ewka.rodek@gmail.com	Poland	Polish Academy of Sciences
Will Rogers	will.rogers@oup.com	United Kingdom	Oxford University Press
Ines Röhrer	ines.roehrer@badw.de	Germany	Bavarian Academy of Sciences and Humanities
Michael Rosner	mike.rosner@um.edu.mt	Malta	University of Malta
Benjamin Rouxel	brouxel@lerobert.com	France	Éditions Le Robert
Manana Ruseishvili-Cartledge	manana.ruseishvili@tsu.ge	Georgia	Ivane Javakhishvili Tbilisi State University
Andriy Rysin	arysin@gmail.com	United States	Independent Researcher
Heete Sahkai	Heete.Sahkai@eki.ee	Estonia	Institute of the Estonian Language
Ana Salgado	anacastrosalgado@gmail.com	Portugal	ILLLP / ACL & FLUP
Guiomar Salvat	guiomarailuo.salvat01@estudiant.upf.edu	Spain	Pompeu Fabra University
Miguel Sánchez Ibáñez	miguelsanchez@uva.es	Spain	University of Valladolid
María Teresa Sanmarco Bande	mt.sanmarco@usc.es	Spain	Universidade de Santiago de Compostela
Antonio San Martín	antonio.san.martin.pizarro@uqtr.ca	Canada	University of Quebec in Trois-Rivières

Author name	Email	Country	Affiliation
Davide Saponaro	davide.saponaro@unint.eu	Italy	Fondazione Rut
Randy Christian Saputra	a2025135389@campus.fcsh.unl.pt	Portugal	NOVA FCSH
Barbara Schmidt-Thieme	bschmidt-thieme@imai.uni-hildesheim.de	Germany	Stiftung Universität Hildesheim
Saghar Sharifi	saghar_sharifi@yahoo.com	Iran	Islamic Azad University Karaj
Jóhannes B. Sigtryggsson	j.sigtryggs@gmail.com, johans@hi.is	Iceland	The Árni Magnússon Institute for Icelandic Studies
Hindrik Sijens	hsijens@fryske-akademy.nl	Netherlands	Fryske Akademy
Raquel Silva	raq.asilva@gmail.com	Portugal	NOVA University
Athina Sioupi	sioupi@del.auth.gr	Greece	Aristotle University of Thessaloniki
Hana Skoumalová	hana.skoumalova@ff.cuni.cz	Czechia	Charles University
Deepak Solanki	deepsoul.solanki@gmail.com	India	Jawaharlal Nehru University, New Delhi
Nicolai Hartvig Sørensen	nhs@dsl.dk	Denmark	Det Danske Sprog- og Litteraturselskab
Michael Spagnol	michael.spagnol@um.edu.mt	Malta	L-Università ta' Malta
Dragana Špica	dspica@unipu.hr	Croatia	Juraj Dobrila University of Pula
Stefania Spina	stefania.spina@unistrapg.it	Italy	Università per Stranieri di Perugia
Irena Srdanović	isrdanovic@unipu.hr	Croatia	Juraj Dobrila University of Pula
Madara Stāde	stade.madara@gmail.com	Latvia	University of Latvia
Ranka Stanković	ranka.stankovic@rgf.bg.ac.rs	Serbia	University of Belgrade
Vasyl Starko	v.starko@ucu.edu.ua	Ukraine	Ukrainian Catholic University
Nika Stefan	nstefan@fryske-akademy.nl	Netherlands	Fryske Akademy
Andrea Stengaard	as@dsl.dk	Denmark	Det Danske Sprog- og Litteraturselskab
Barbora Štěpánková	stepankova@ufal.mff.cuni.cz	Czechia	Charles University
Martin Stluka	martin.stluka@ff.cuni.cz	Czechia	Charles University
Philipp Stöckle	philipp.stoeckle@oeaw.ac.at	Austria	Austrian Academy of Sciences
Petra Storjohann	storjohann@ids-mannheim.de	Germany	Leibniz Institute for the German Language (IDS)
Silga Sviķe	silga.svike@gmail.com	Latvia	Ventspils University of Applied Sciences
Pius ten Hacken	pius.ten-hacken@uibk.ac.at	Austria	Universität Innsbruck
Dušica Terzić	dusica.terzic@fil.bg.ac.rs	Serbia	University of Belgrade

Author name	Email	Country	Affiliation
Carole Tiberius	c.p.a.tiberius@hum.leidenuniv.nl, carole.tiberius@ivdnt.org	Netherlands	Instituut voor de Nederlandse Taal / Leiden UniversityCentre for Linguistics
Svetlana Timoshenko	nyrestein@gmail.com	Russia	Russian Academy of Science
Sergi Torner	sergi.torner@upf.edu	Spain	Pompeu Fabra University
Jesús Torres del Rey	jtorres@usal.es	Spain	Universidad de Salamanca
Ágoston Tóth	toth.agoston@arts.unideb.hu	Hungary	University of Debrecen
Lars Trap-Jensen	ltj@dsl.dk	Denmark	Society for Danish Language and Literature
Catherine Trekker	catherine.trekker-seguin@uqtr.ca	Canada	University of Quebec in Trois-Rivières
Aurora Troncoso Ruiz	aurora.troncosoruiz@ru.nl	Netherlands	Radboud University Nijmegen
Marijana Tuta Dujmović	mtuta@suvag.hr	Croatia	Polyclinic for Hearing and Speech Rehabilitation SUVAG
Maria Tuulik	maria.tuulik@eki.ee	Estonia	Institute of the Estonian Language
Richard Two Dogs	r.twodogs@gmail.com	United States	Lakota Language Consortium
Efi Tzelepi	etzelepi@helit.duth.gr	Greece	Democritus University of Thrace
Ioannis Tziogkas	john.tziog@gmail.com	Greece	Harokopio University of Athens
Thordis Ulfarsdottir	disa@hi.is	Iceland	The Árni Magnússon Institute for Icelandic Studies
Jan Ullrich	jfu@lakhota.org	Czechia	Lakota Language Consortium
Nico Urbach	nico.urbach@uni-marburg.de	Germany	Research Center Deutscher Sprachatlas
Anna Vacalopoulou	avacalop@athenarc.gr	Greece	Institute for Language and Speech Processing / Athena Research Center
Ene Vainik	Ene.Vainik@eki.ee	Estonia	Institute of the Estonian Language
Carlos Valcárcel Riveiro	carlos.valcarcel@uvigo.gal, carlos.valcarcel.riveiro@gmail.com	Spain	Universidade de Vigo
Johan van der Zwaag	jvanderzwaag@fryske-akademy.nl	Netherlands	Fryske Akademy
Silver Vapper	silver.vapper@eki.ee	Estonia	Institute of the Estonian Language
Iuliia Verbynenko	yulia_verbinenko@yahoo.co.uk	Ukraine	Ukrainian Lingua-Information Foundation of NAS of Ukraine

Author name	Email	Country	Affiliation
Federica Vezzani	federica.vezzani@unipd.it	Italy	Università degli Studi di Padova
Giuliana Elizabeth Vilela Ruiz	giulianaelizabeth.vilelaruiz@ilc.cnr.it	Italy	Institute Of Computational Linguistics “A. Zampolli”
Federico Villa	federico.villa03@icatt.it	Italy	Independent Researcher
Martina Vokáčová	martina.vokacova@ff.cuni.cz	Czechia	Charles University
Perina Vukša Nahod	pvuksa@ihjj.hr	Croatia	Institute of Croatian Language
Zongyuan Wang	z.wang1@unior.it	Italy	Università degli Studi di Napoli L'Orientale
Kate Wild	kate.wild@oup.com	United Kingdom	Oxford University Press
Tarrin Wills	tarrin@hum.ku.dk	Denmark	University of Copenhagen
Jonas Winnerlöv	jonas.winnerlov@europarl.europa.eu	Sweden	European Parliament
Sylwia Wojciechowska	sylwiab@amu.edu.pl	Poland	Adam Mickiewicz University
Yingtian Wu	wyt15858810001@163.com	Czechia	Lexical Computing
Mykyta Yablochkov	gezartos@gmail.com	Ukraine	Ukrainian Lingua-Information Foundation of NAS of Ukraine
Shigeru Yamada	shyamada@waseda.jp	Japan	Waseda University
Yelena Yerznkyan	yerznkyan@ysu.am	Armenia	Yerevan State University
Akerke Yessenali	akerke.yessenali@univ-lorraine.fr	France	Université de Lorraine
Patrick Zabalbeascoa	patrick.zabalbeascoa@upf.edu	Spain	Pompeu Fabra University
Manjola Zaçellari	manjolazacellari@uamd.edu.al	Albania	Aleksander Moisiu University
Fabio Zanda	fabio.zanda@uniroma1.it	Italy	Sapienza University of Rome
Marija Žarković Eriksson	marija.zarkovic@autonoma.cat	Spain	Universitat Autònoma de Barcelona
Alexander Ziem	ziem@hhu.de	Germany	Heinrich Heine University Düsseldorf
Tanara Zingano Kuhn	tanarazingano@outlook.com	Portugal	University of Coimbra
Valeria Zotti	valeria.zotti@unibo.it	Italy	Alma Mater Studiorum Università di Bologna

INDEX OF AUTHORS

A

Abdelzaher, Esra 267
Abel, Andrea 183
Afentoulidou, Vassiliki 8
Akinremi, Luke Omoyemi 184
Alamdar, Abolfazl 253
Alonso, Mikel 11
Ammari, Maria 14
An, Jinsan 78
Arčon, Tjaša 228
Arias-Arias, Iván 185, 272
Arias-Badia, Blanca 142

B

Bago, Petra 186
Bajčić, Martina 187
Bandini, Michela 17
Barbosa, Sílvia 146
Baumann, Andreas 4
Beliga, Slobodan 208
Bellandi, Andrea 17, 101, 138
Bernal, Elisenda 142
Berthemet, Elena 188
Bichlmeier, Harald 189
Billero, Riccardo 190
Black Bear Jr., Ben 191
Blahuš, Marek 192, 229
Böbel, Nina 282
Boguslavsky, Igor 193
Botezat, Onorina 81
Bothma, Theo J.D. 217
Božinovski, Biljana 20
Brač, Ivana 195
Brankačkec, Katja 196
Breathnach, Cormac 197
Brescia-Zapata, Marta 142
Brůhová, Gabriela 23
Burkart, Rukiye Smilla 198

C

Camilleri, John J. 257
Cane, Edmond 25
Caruso, Valeria 199, 274
Castaldi, Elena 27
Chadjipapa, Elina 29
Chen, Lian 200, 285
Chew III, John J. 32
Chirino, Dalí 264
Christofidou, Anastasia 8
Cimander, Luisa 34, 201
Colman, Lut 264
Cui, Li 202
Cvikić, Lidija 133
Cvrček, Václav 36
Czerski, Dariusz 125
Czyż, Tomasz 114

D

Dao, Huy-Linh 200
de Does, Jesse 67
De Meo, Anna 274
de Oliveira Maia-Pires, Flávia 203
de Schryver, Gilles-Maurice 58, 61, 204, 205
De Tier, Veronique 38
Deilke, Janina 109
Del Rey Adanalian, Clara Anush 142
Delaporte, Alexander 200
Depuydt, Katrien 38, 67
Despot, Kristina Š. 42
Dikonov, Vyacheslav 193
Dinić Marinković, Milica 236
Domínguez Vázquez, María José 185, 272
Dorozhynska, Alona 127
Dosi, Ifigeneia 29
Dringó-Horváth, Ida 206
Ducceschi, Luca 183

E

Eder, Elisabeth 227
Ellul, Dwayne 257
Elsner, Daniel 227
Eterović, Ivana 186

F

Fannee, Mathieu 38
Ferrett, Emma 32, 45
Fiedler, Sophia 207
Filipović Petrović, Ivana 208
Fioravanti, Irene 155
Fiotaki, Alexandra 105
Fischer, Hanna 209

Flörke, Ida 242
Fomat, Emmanuel Sylvain 47
Franceschi, Daniele 51
Franjić, Mihaela 20
Frankenberg-Garcia, Ana 118, 142
Freixa, Judit 70
Freixas, Margarita 210
Frolova, Tatiana 193
Fuentes Morán, M.ª Teresa 265

G

Gabrovšek, Dejan 135
Ganslmayer, Christine 211
Gantar, Polona 85, 228
Garcia Casado, Marta 142, 212
Garriga Martínez, Pol 213
Gärtig-Bressan, Anne-Kathrin 53
Gavriilidou, Zoe 29, 162, 214, 215
Geyken, Alexander 216
Ghorvei, Amir 45
Giannakou, Aretousa 8
Giouli, Voula 282
Giovannetti, Emiliano 17
Glebus, Marius 56
Gobel, Peter 20
Gouws, Rufus H. 217
Gramellini, Rita 179
Grasmanis, Mikus 241

H

Hajič, Jan 6
Han, Wanjing 58, 61
Heid, Ulrich 149
Heinisch, Barbara 218
Heinonen, Tarja 70
Herman, Ondřej 219
Hernández, Esther 210
Heylen, Kris 67, 70, 130, 280
Holmer, Louise 73
Horvat, Marijana 75
Horvat, Martina 75
Hudeček, Lana 233
Humbley, John 81

I

Iaroshenko, Polina 220
Igareda González, Paula 142
Inshakova, Evgeniya 193

J

Jakubíček, Miloš 108, 176, 192, 229
Janda, Laura A. 230
Ježovnik, Janoš 89, 135
Jiménez Casas, Daniel 142
Johannsson, Ellert Thor 221, 269, 275
Jozić, Željko 75
Jung, Hae-Yun 70, 78, 202

K

Kabashi, Besim 285
Kabatnik, Susanne 198
Kallas, Jelena 271
Kano, Makimi 20
Kern, Boris 222
Kernerman, Ilan 70, 81, 130, 280
Kettnerová, Václava 223, 225
Kiil, Ahto 271
Kim, Agnes 224
Klemen, Matej 42
Kolářová, Veronika 223, 225
Konstantinidou, Evanthia 215
Koppel, Kristina 160, 249
Koppensteiner, Wolfgang 227
Kopylova, Mariia 226
Korecky-Kröll, Katharina 227
Kosem, Iztok 85, 228
Kovář, Vojtěch 108, 176, 229
Kovařík, František 229
Kováříková, Dominika 230
Krawczyk, Iwona 243
Krek, Simon 85, 228
Kroon, Martin 264
Krstev, Cvetana 236
Kruse, Theresa 149
Krvina, Domen 89
Kryńska, Katarzyna 95
Kugic, Amila 87

L

Laluashvili, Tamar 234
Lameli, Alfred 209
Langemets, Margit 160, 249
Lazursky, Alexandre 193
Ledinek, Nina 89
Lee, Jun 78, 202
Lee, Seulki 92
Lew, Robert 2, 142

Lewandowska-Tomaszczyk, Barbara 231
Licurgo Pires, Silvio 203
Lindemann, David 81, 232
Lombardi, Pia Carmela 53
Lorentzen, Henrik 70, 268
Loukachevitch, Natalia 220
Lovrić Jović, Ivana 233
Lutsepp, Liina 160

M

Maask, Hanna 160, 249
Majdak, Magdalena 95
Makhatadze, Marine 239
Margalitadze, Tinatin 234
Marjanović, Saša 263
Marklová, Anna 273
Marković, Aleksandra 236
Marszałek, Jagoda 95, 243
Martín-Cancela, Elena 184
Martins, Leonor 98
Masbah, Abdelouahab 101
Mathioudakis, Nikos 105
Matuška, Ondřej 108, 284
Mchedlishvili, Ketevan 234
Mederake, Nathalie 209
Meißner, Cordula 109
Meladze, Giorgi 234
Menagarishvili, Olga 112
Mertens, Koen 38
Meyer, Peter 237
Michelizza, Mija 89
Michta, Tomasz 114, 116, 118
Middell, Gregor 216
Mihaljević, Ana 238
Mihaljević, Josip 238
Mika, Dorota 244
Mikkelsen, Agnes Aggergaard 242
Milà Garcia, Alba 142
Mirković-Birtašić, Dejana 263
Mírovský, Jiří 223, 225
Mitsiaki, Maria 29
Moore, Julie 120
Movsisyan, Diana 277
M. Pintér, Tibor 239
Mroczyńska, Katarzyna 116

N

Nahod, Bruno 240
Nam, Kilim 70, 78, 202
Namkung, Seol 202
Nešpore-Bērzkalne, Gunta 241
Nicholson, Phoebe 123
Nied Curcio, Martina 211
Nimb, Sanni 242
Norman, Nathalie Hau 242
Nouvel, Damien 200
Nowak, Krzysztof 243, 244

O

Ogrodniczuk, Maciej 125
Okropiridze, Giorgi 234
Ölander Aadland, Emma Josefin 245
Olbrich, Michal 225
Ó Mianáin, Pádraig 197
Ostapova, Iryna 127
Ostroški Anić, Ana 42, 130, 133, 187

P

Papanagiotou, Christos 8
Paulsen, Geda 271
Pavić, Martina 133
Perak, Benedikt 246
Perdih, Andrej 89, 135
Petreková, Katarína 192
Petric Žižić, Špela 89
Piccini, Silvia 17, 101, 138
Pivoňková, Klára 36
P. Márkus, Katalin 206, 239
Pődör, Dóra 239
Prangel, Esta 160, 249
Prossliner, Linda 247

Q

R

Raaf, Manuel 251
Randermann, Anna-Lena 109
Raušová, Veronika 23
Rebosio, Laura 248
Rees, Geraint Paul 118, 142, 199
Riehl, Anastasia 32, 45
Risberg, Lydia 160, 249
Rodek, Ewa 95
Rogers, Will 123
Röhrer, Ines 251
Rosner, Michael 81
Rouxel, Benjamin 70

Rusieshvili-Cartledge, Manana 239
Rysin, Andriy 258

S

Sahkai, Heete 271
Salgado, Ana 5, 70, 81, 98
Salvat, Guiomar 142
Sánchez Ibáñez, Miguel 70
Sanmarco Bande, María Teresa 185, 272
San Martín, Antonio 252
Saponaro, Davide 138
Saputra, Randy Christian 146
Schmidt-Thieme, Barbara 149
Sharifi, Saghar 253
Sigtryggsson, Jóhannes B. 221, 254
Sijens, Hindrik 259
Silva, Raquel 81
Sioupi, Athina 282
Skoumalová, Hana 255
Solanki, Deepak 256
Sørensen, Nicolai Hartvig 151, 242
Spagnol, Michael 257
Špica, Dragana 246
Spina, Stefania 155
Srdanović, Irena 246
Stāde, Madara 241
Stanković, Ranka 236
Starko, Vasyl 258
Stefan, Nika 259
Stengaard, Andrea 151
Štěpánková, Barbora 196
Stluka, Martin 36
Stöckle, Philipp 227
Storjohann, Petra 34, 70, 158, 260
Sviķe, Silga 261

T

ten Hacken, Pius 262
Terzić, Dušica 263
Tiberius, Carole 67, 264
Timoshenko, Svetlana 193
Torner, Sergi 142
Torres del Rey, Jesús 265
Tóth, Ágoston 266, 267
Trap-Jensen, Lars 70, 268
Trekker, Catherine 252
Troncoso Ruiz, Aurora 142
Tuta Dujmovic, Marijana 240
Tuulik, Maria 160, 249
Two Dogs, Richard 191
Tzelepi, Efi 162
Tziogkas, Ioannis 8

U

Ulfarsdottir, Thordis 221, 269
Ullrich, Jan 191
Urbach, Nico 209

V

Vacalopoulou, Anna 81, 270
Vainik, Ene 160, 271
Valcárcel Riveiro, Carlos 185, 272
van der Zwaag, Johan 259
Vapper, Silver 249
Verbynenko, Iuliia 127
Vezzani, Federica 81, 130
Vilela Ruiz, Giuliana Elizabeth 138
Villa, Federico 183
Vokáčová, Martina 273
Vukša Nahod, Perina 165

W

Wang, Zongyuan 274
Wild, Kate 3, 123
Wills, Tarrin 275
Winnerlöv, Jonas 167
Wojciechowska, Sylwia 169
Wu, Yingtian 192

X

Y

Yablochkov, Mykyta 127
Yamada, Shigeru 276
Yerznkyan, Yelena 277
Yessenali, Akerke 171

Z

Zabalbeascoa, Patrick 142
Zaçellari, Manjola 174
Zanda, Fabio 155
Žarković Eriksson, Marija 210, 278
Ziem, Alexander 282
Zingano Kuhn, Tanara 176
Zotti, Valeria 179