Electronic lexicography in the 21st century (eLex 2025)

Book of abstracts

edited by Iztok Kosem
Miloš Jakubíček
Marek Medveď
Karolina Zgaga
Špela Arhar Holdt
Tina Munda
Ana Salgado

18-20 November 2025

elex.link/elex2025

Electronic lexicography in the 21st century (eLex 2025)

Book of abstracts

Edited by
Iztok Kosem
Miloš Jakubíček
Marek Medveď
Karolina Zgaga
Špela Arhar Holdt
Tina Munda
Ana Salgado

Issued by
Centre for Language Resources and Technologies
University of Ljubljana
Ljubljana University Press
Faculty of Arts

For the issuer
Mojca Schlamberger Brezar
Dean of the Faculty of Arts
University of Ljubljana

Published by
Ljubljana University Press
Faculty of Arts

For the publisher
Gregor Majdič
Rector of the University of Ljubljana

November 2025, Bled, Slovenia
elex.link/elex2025

eLex 2025

ORGANIZERS

UNIVERZA V LJUBLJANI
University of Ljubljana

cjvt Centre for Language Resources and Technologies

SPONSORS

Hornby
A. S. Hornby Educational Trust

Cambridge Dictionary

LEXICAL COMPUTING

elex.link/elex2025

Organizing Committee

Iztok Kosem
Simon Krek
Polona Gantar
Špela Arhar Holdt
Tina Munda
Sara Kos
Tinca Lukan

Scientific committee

Andrea Abel
Špela Arhar Holdt
Lut Colman
Ivana Filipović Petrović
Polona Gantar
Yongwei Gao
Radovan Garabík
Alexander Geyken
Vojko Gorjanc
Kris Heylen
Louise Holmer
Miloš Jakubíček
Madis Jürviste
Jelena Kallas
Ilan Kernerman
Annette Klosa-Kückelhaus
Svetla Koeva
Kristina Koppel
Iztok Kosem
Vojtěch Kovář
Simon Krek
Margit Langemets
Lothar Lemnitzer
Pilar León Araúz
Robert Lew
Veronika Lipp
Nikola Ljubešić
Henrik Lorentzen
Michal Měchura
Tomasz Michta
Monica Monachini
Tina Munda
Hilary Nesi
Lionel Nicolas
Sanni Nimb
Sussi Olsen
Vincent Ooi
Marco Passarotti
Laurent Romary
Ana Salgado
Hindrik Sijens
Emma Sköldberg
Ranka Stanković
Egon W. Stemle
Kristina Strkalj Despot
Arvi Tavast
Carole Tiberius
Yukio Tono
Lars Trap-Jensen
Sascha Wolfer
Tanara Zingano Kuhn

Contents

Book of Abstracts

KEYNOTE TALKS

Large language models for lexicography

Marko Robnik-Šikonja

Faculty of Computer and Information Science, University of Ljubljana, Slovenia
E-mail: marko.robniksikonja@fri.uni-lj.si

Abstract

Currently, large language models (LLMs) are redefining methodological approaches in many scientific areas, including linguistics and lexicography. LLMs are pretrained on huge text corpora by predicting the next tokens and adapted for human interaction with the instruction following datasets. This does not make them immune to hallucinations and biases, requiring a human-in-the-loop approach. In the context of lexicography, LLMs can be used to support several tasks. We will present how the information contained in language databases can be utilized to improve LLMs on lexicographic tasks. Our current methodology is based on knowledge graph extraction, continued pretraining of LLMs, prompt engineering, and semi-automatic evaluation.

LLMs and Lexicography at the Dutch Language Institute

Carole Tiberius[1,2], Jesse de Does[2]

[1] Leiden University Centre for Linguistics, Netherlands
[2] Dutch Language Institute, Netherlands
E-mail: c.p.a.tiberius@hum.leidenuniv.nl, jesse.dedoes@ivdnt.org

Abstract

The Dutch Language Institute (INT) has a long tradition compiling historic and contemporary dictionaries and other types of lexicographic databases, mainly for Dutch but also for some other languages with a relation to Dutch. Lexicographic work at the institute is computer-supported but there is still a great deal of manual work involved. Therefore, INT is exploring how new technologies (including LLMs) can be used for optimising different parts of the lexicographic work without compromising data quality and reliability. After a brief overview of various pilot studies conducted at the institute, we will take a closer look at how we can make the implementation of Hanks' Corpus Pattern Analysis procedure (as it is used in the context of the project Woordcombinaties) more intelligent. This way, we hope to ultimately realise Patrick Hanks' vision that "it seems likely that a large part of the work that is currently being carried out by hand will be automated in the not-too-distant future" (Hanks 2013;247).

We need to talk about data structures in lexicography

Michal Měchura

Lexical Computing and Dublin City University
E-mail: michmech@mail.muni.cz

Abstract

It has been almost half a century since we started "doing" lexicography on computers. Let's stop for a minute now and take a critical look at the data models we have been using to represent the structure of dictionaries in dictionary writing systems and other software.

In this talk, I will trace the history of lexicographic data modelling from its beginnings as text markup for retro-digitised dictionaries, to the present day when most dictionaries are born-digital. I will show that, regardless of which notation we use (XML, JSON or other), the underlying design pattern is almost always a tree structure in which the various content items (headwords, senses, definitions…) are arranged in a parent-child hierarchy.

I will argue that the tree-structured pattern is not expressive enough to handle some phenomena that occur in dictionaries, such as entry-to-entry cross-references, the placement of multiword subentries, and complex hierarchies of subsenses. These things would be easier to manage in a graph-based data structure, such as a relational database or a Semantic Web-style knowledge graph.

Dictionary projects which insist on a purely tree-structured data model are failing to make full use of the digital medium. But upgrading to a graph-based data model is difficult because tree-structured thinking is entrenched in the minds of lexicographers and dictionary users alike. This talk will conclude with an introduction to DMLex, a recently standardised "Data Model for Lexicography" which aims to ease this transition by being a hybrid model, combining tree structures where possible with graph structures where necessary.

PRESENTATIONS

The lemma dilemma, Slovene version

Polona Gantar[1], Cyprian Laskowski[2], Simon Krek[2,3]

[1] Faculty of Arts, University of Ljubljana
[2] Centre for Language Resources and Technologies, University of Ljubljana
[3] Jožef Stefan Institute
E-mail: apolonija.gantar@guest.arnes.si, cyp@cjvt.si, simon.krek@guest.arnes.si

Abstract

In lexicography, one of the long-standing issues is understanding the nature of its core element of description commonly referred to as the headword (in DMLex and traditional lexicography), canonical form (in OntoLex and the Lexical Markup Framework – LMF), orthographic form (in the Text Encoding Initiative – TEI Lex0), lemma (in Wikidata), or lexical unit. With the transition from paper to digital environments, both the nature of this element and its description have evolved. At the heart of the "lemma dilemma" lies the relationship between form (particularly in logographic writing systems) and sense—the (description of a) concept intended to be meaningful to humans.

In this paper, we describe how the headword/lemma phenomenon is addressed in the Digital Dictionary Database for Slovene (DDDS). The DDDS includes two types of lexical units: concepts and named entities. The latter are defined lexicographically in the same manner as concepts and are included in the DDDS due to the need to provide information on inflection, pronunciation, normative status, or other linguistic factors.

Lexical units are mechanically divided into single lexeme units and multiword expressions (MWEs), based on their single-word or multi-word status in the Slovene writing system. Typologically, MWEs (excluding multiword named entities) are further divided into compounds and phrases.

The ultimate goal of the DDDS is to compile all types of information about the Slovene lexicon in a single database with a unified data model. Like other Slavic languages, Slovene has a very rich morphology, which often presents a dilemma for lexicographers when choosing the most appropriate word form to represent a concept—i.e., the headword. The DDDS includes a vast number of word forms with morphological data, including pronunciation and stress. Currently, this number stands at 9,312,865.

In the data model, a collection of morphologically linked word forms is defined as a LEXEME. According to this principle, a typical Slovene noun (associated with a unique LEXEME ID) includes 18 word forms, combining three grammatical numbers (singular, dual, plural) and six grammatical cases (nominative, genitive, dative, accusative, locative, instrumental).

As of now, the DDDS contains 395,613 lexemes. When forming a LEXICAL UNIT—which adds the conceptual or semantic layer of description—one word form must be selected to represent the lexical unit. This selected form is traditionally considered the headword, canonical form, or lemma. Consequently, the same LEXEME ID can be used for multiple LEXICAL UNITS, even if different word forms serve as the "headword" for each.

A practical example of this situation is a singular–plural noun pair where the same LEXEME ID and two different word forms are used as headwords to define two distinct concepts: "jajce" (Eng. egg, nominative singular) and "jajca" (Eng. testicles, nominative plural).

In the paper, we will provide a more detailed explanation of these principles, supported by additional examples.

Keywords: digital dictionary database; Slovene language; Slovene morphology;

headword status; standards in lexicography

References

Gorjanc, V., Gantar, P., Kosem, I. & Krek, S. (eds). (2017). *Dictionary of modern Slovene: problems and solutions*. 1st ed., e-ed. Ljubljana: Ljubljana University Press, Faculty of Arts. ilustr. Book series Prevodoslovje in uporabno jezikoslovje. ISBN 978-961-237-913-1, ISBN 978-961-237-914-8.

Gantar, P. (2020). Dictionary of modern Slovene: from Slovene lexical database to digital dictionary database. *Rasprave Instituta za hrvatski jezik i jezikoslovlje*, 46(2), pp. 589–602, ilustr. ISSN 1849-0379. Available at: https://hrcak.srce.hr/index.php?show=clanak&id_clanak_jezik=356570, DOI: 10.31724/rihjj.46.2.7.

Lexicography and Generative Artificial Intelligence for contextualised meaning

Theo J.D. Bothma[1], Rufus H. Gouws[2]

[1] Department of Information Science, University of Pretoria, South Africa
[2] Department of Afrikaans and Dutch, Stellenbosch University, South Africa
E-mail: theo.bothma@up.ac.za, rhg@sun.ac.za

Abstract

The focus of this paper is on Generative Artificial Intelligence (GenAI), chatbots and some implications for lexicography and dictionary use. It has been well documented that chatbots originally tended to "hallucinate" if they did not have an answer to the prompt put to them. Much larger training databases have, however, been developed and chatbots have become more accurate. Multiple iterations of chatbots from a variety of companies have been released, including specialised chatbots for different environments. AI and chatbots have also been frequent topics in recent lexicographic research and have been employed in dictionary compilation and the preparation of writing assistants (cf., e.g., Li et al. (2023), De Schryver (2023), Fuertes-Olivera (2024), Lew 2024 & Li & Tarp (2025)). From a lexicographic perspective, the importance of linking between dictionaries and other information tools (cf., e.g., Bothma and Gouws 2022, Bothma and Fourie 2024, Bothma and Fourie 2025) also becomes relevant for lexicographic uses of chatbots.

The use of GenAI as an information tool to provide information to end-users (readers) who have a specific information need when reading a text, i.e., a text reception information need, is discussed in detail. It has been shown that GenAI can provide content similar to a dictionary, but that it cannot provide contextualised answers, i.e., the reader is still dependent on their own evaluation of the GenAI-provided content to determine the meaning of the word or phrase in context. If sufficient context is provided in the prompt, the chatbot often provides only a single meaning / sense. If the chatbot misunderstood the context provided in the prompt, it could easily provide an incorrect meaning. If then queried (through a follow-up prompt) why it chose a specific meaning, it could not provide any explanation. Quite recently, however, this changed, and most chatbots now have two modes, a "search" mode and a "thinking / reasoning mode", i.e., it is able to argue logically about its different proposed meanings in context and tends to offer a solution. This feature is discussed at the hand of a number of examples containing specific keywords that determine the correct interpretation in context, as well as examples with potentially ambiguous part-of-speech and syntactic analyses, using two different chatbots, viz. ChatGPT o3-mini and DeepSeek-V3 (DeepThink-R1). Based on the limited number of examples, it seems as if the chatbots can provide

correct contextual meaning and logically motivate the choice of meaning in context, based on their critical analysis and thinking skills, typically associated with humans. Unfortunately, however, it still "hallucinates" if it has no answer, as will be shown from one non-lexicographic example, and the reader remains responsible to critically evaluate any GenAI responses – "lector caveat." Nevertheless, in slightly more than two years, tremendous progress has been made, and one can only speculate what next developments would be.

These developments raise the question of what the role of dictionaries and the role of lexicographers will be in future in an AI-enhanced world. In conclusion, a few suggestions will be offered about lexicographic databases, appropriate interfaces, access to additional lexicographic and non-lexicographic data, refining dictionary definitions, multifunctional dictionaries, and the reuse of lexicographic information in different applications. The traditional role of dictionaries to document the status and history of a language is still a very important function and needs to be encouraged, especially in environments with limited language resources. However, exploring new commercial ventures, incorporating latest technologies, would be essential to the future of the discipline and industry.

Keywords: Generative Artificial Intelligence; Chatbots; Text reception; Dictionary consultation; Contextualised meaning

References

Bothma, T.J.D. & Fourie, I. (2024). Enhancing Conceptualisations of Information Behaviour Contexts through Insights from Research on E-Dictionaries and E-Lexicography. *Information Research an international electronic journal*, 29(2), pp. 179–197. Available at: https://doi.org/10.47989/ir292821.

Bothma, T.J.D. & and Fourie, I. (2025). Contextualised Dictionary Literacy, Information Literacy, and Information Behaviour In the E-environment. *Library Management*, 46(1/2), pp. 14–28. Available at: https://10.1108/LM-08-2023-0082.

Bothma, T.J.D. & Gouws, R.H. (2022). Information Needs and Contextualization in the Consultation Process of Dictionaries that are Linked to E-texts. *Lexikos*, 32(2), pp. 53–81. Available at: https://doi.org/10.5788/32-2-1697.

De Schryver, G.-M. (2023). Generative AI and Lexicography: The Current State of the Art Using ChatGPT. *International Journal of Lexicography*, 36(4), pp. 355–387. Available at: https://doi.org/10.1093/ijl/ecad021.

Fuertes-Olivera, P.A. (2024). Making Lexicography Sustainable: Using AI and Reusing Data for Lexicographic Purposes. *Lexikos*, 34, pp. 123–140. Available at: https://orcid.org/0000-0003-3831-5377.

Huete-García, Á. & Tarp, S. (2024). Training an AI-based Writing Assistant for Spanish Learners: The Usefulness of Chatbots and the Indispensability of Human-assisted

Intelligence. *Lexikos*, 34, pp. 21–40. Available at: https://doi.org/ 10.5788/34-1-1862.

Lew, R. (2024). Dictionaries and Lexicography in the AI Era Linguistics. *Humanities and Social Sciences Communications*, 11(1), pp. 1–8. Available at: https://doi.org/ 10.1057/s41599-024-02889-7.

Li, J., Ren, X., Jiang, X. & Chen, C.-H. (2023). Exploring the Use of ChatGPT in Chinese Language Classrooms. *International Journal of Chinese Language Teaching*, 4(3), pp. 36–55. Available at: https://doi.org/10.46451/ijclt.20230303.

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

The role of subjectivity in lexicography: Experiments towards data-driven labeling of informality

Lydia Risberg[1, 2], Eleri Aedmaa[1], Maria Tuulik[1], Margit Langemets[1], Ene Vainik[1], Esta Prangel[1], Kristina Koppel[1], Hanna Pook[1]

[1] Institute of the Estonian Language, Roosikrantsi 6, Tallinn, Estonia
[2] University of Tartu, Jakobi 2, Tartu, Estonia
E-mail: lydia.risberg@eki.ee, eleri.aedmaa@eki.ee, maria.tuulik@eki.ee, margit.langemets@eki.ee, ene.vainik@eki.ee, esta.prangel@eki.ee, kristina.koppel@eki.ee, hanna.pook@eki.ee

Abstract

Language corpora have long been used in linguistics and lexicography, but recent developments now allow large language models (LLMs) to support or even transform these fields. This study investigates the potential of LLMs for annotating informal language use in Estonian – a language underrepresented in LLM training data yet supported by a large corpus. Focusing on the informal register label used in the Dictionary of Standard Estonian, we explore whether LLMs can assist lexicographers in determining the informal label. This paper describes two experiments that make use of LLMs, including GPT, Gemini, and Claude. The first experiment yielded useful insights but also highlighted necessary improvements. In the second experiment, we evaluated the LLMs' consistency and accuracy in categorizing words as informal or neutral/formal. Results showed that LLMs achieved around 76% agreement with expert human annotators, significantly above random chance, suggesting their usefulness as a supplementary resource in lexicography. GPT-4o demonstrated high accuracy, stability, and cost-efficiency, making it a reliable candidate for such a lexicographic task. The study highlights the inherent subjectivity in register labeling and the value of combining corpus data, expert judgment, and LLM output. Overall, LLMs represent a promising tool for modern dictionary work.

Keywords: large language models; register labels; Estonian; lexicography; informal language

References

Anthropic. (2024). Claude [Large Language Model]. Accessed at: https://www.anthropic.com/claude. (16 September 2025)

Baayen, R.H. (2024). The wompom. *Corpus Linguistics and Linguistic Theory*, 20(3), pp. 615–648.

Beal, J., Lukač, M. & Straaijer, R. (2023). *The Routledge Handbook of Linguistic Prescriptivism.* London, New York: 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 *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. New York, NY, USA: Association for Computing Machinery, pp. 610–623.

Bi, X., Chen, D., Chen, G., Chen, S., Dai, D., Deng, C., Ding, H., Dong, K., Du, Q., Fu, Z. & Gao, H. et al. (2024). DeepSeek LLM: Scaling open-source language models with longtermism. Available at: https://doi.org/10.48550/arXiv.2401.02954. (16 September 2025)

Biber, D. & Conrad, S. (2009). *Register, Genre, and Style (Cambridge Textbooks in Linguistics)*. Cambridge: Cambridge University Press.

Biber, D. & Egbert, J. (2023). What is a register? Accounting for linguistic and situational variation within – and outside of – textual varieties. *Register Studies*, 5. Available at: https://www.jbe-platform.com/content/journals/10.1075/rs.00004.bib. (16 September 2025)

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A. et al. (2020). Language Models are Few-Shot Learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan & H. Lin (eds.) *Advances in Neural Information Processing Systems*, 33, pp. 1877–1901. Curran Associates, Inc.

Claude Sonnet 3 = Anthropic (2024). The Claude 3 Model Family: Opus, Sonnet, Haiku. Accessed at: https://www.anthropic.com/news/claude-3-family. (16 September 2025)

Claude Sonnet 3.5 = Anthorpic. (2024). Claude 3.5 Sonnet. Accessed at : https://www.anthropic.com/news/claude-3-5-sonnet. (16 September 2025)

Claude Sonnet 3.7 = Anthropic. (2025). Claude 3.7 Sonnet and Claude Code. Accessed at : https://www.anthropic.com/news/claude-3-7-sonnet. (16 September 2025)

Claude Sonnet 4 = Anthorpic. (2025). Introducing Claude 4. Accessed at : https://www.anthropic.com/news/claude-4. (16 September 2025)

Claude Opus 4 = Anthorpic. (2025). Introducing Claude 4. Accessed at: https://www.anthropic.com/news/claude-4. (16 September 2025)

Claude Opus 4.1 = Anthropic. (2025). System Card Addendum: Claude Opus 4.1. Accessed at: https://www.anthropic.com/claude-opus-4-1-system-card. (16 September 2025)

Davies, M. (2025a). Comparing and Integrating Information from Corpora and AI/LLMS. *EURALEX Talks*, 16 April 2025. Accessed at: https://videolectures.net/videos/EURALEXTalks_davies_information. (16 September 2025)

Davies, M. (2025b). Integrating AI / LLMs into English-Corpora.org. Accessed at: https://www.english-corpora.org/ai-llms. (16 September 2025)

De Schryver, G.-M. (2023). Generative AI and Lexicography: The Current State of the Art Using ChatGPT. *International Journal of Lexicography*, 36(4), pp. 355–387.

DSE 2018 = Eesti õigekeelsussõnaraamat ÕS 2018 [Dictionary of Standard Estonian]. T. Erelt, T. Leemets, S. Mäearu & M. Raadik (eds.). Eesti Keele Instituut. Tallinn: EKSA.

EKI Combined Dictionary = *Eesti Keele Instituudi ühendsõnastik 2025.* Eesti Keele Instituut, Sõnaveeb. Accessed at: https://sonaveeb.ee. (16 September 2025)

EuroLLM = Martins, P.H., Alves, J., Fernandes, P., Guerreiro, N.M., Rei, R., Farajian, A., Klimaszewski, M., Alves, D.M., Pombal, J., Faysse, M. and Colombo, P. (2025) *EuroLLM-9B: Technical Report.* Available at: https://doi.org/10.48550/ arXiv.2506.04079. (16 September 2025)

Gemini Team, Google. (2023). Gemini: A Family of Highly Capable Multimodal Models. Available at: https://doi.org/10.48550/arXiv.2312.11805. (16 September 2025)

Gemini 1.5 Pro = Gemini Team, Google. (2024). *Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context.* Available at: https://doi.org/10.48550/arXiv.2403.05530. (16 September 2025)

GPT-4 = Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.-L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S. et al. (2023). *Gpt-4 technical report.* Available at: https://doi.org/10.48550/arXiv.2303.08774. (16 September 2025)

GPT-4.1 = OpenAI. (2025). Introducing GPT-4.1 in the API. Accessed at: https://openai.com/index/gpt-4-1/. (16 September 2025)

GPT-4.5 = OpenAI. (2025). OpenAI GPT-4.5 System Card. Accessed at: https://openai.com/index/gpt-4-5-system-card/. (16 September 2025)

GPT-4o = Hurst, A., Lerer, A., Goucher, A.-P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A.J. & et al. (2024). GPT-4o system card. Available at: https://doi.org/10.48550/arXiv:2410.21276. (16 September 2025)

GPT-5 = OpenAI. (2025). GPT-5 System Card. Accessed at: https://cdn.openai.com/gpt-5-system-card.pdf. (16 September 2025)

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., et al. (2024). *The llama 3 herd of models.* Available at: https://doi.org/10.48550/arXiv.2407.21783. (16 September 2025)

Grok 3 =. xAI. (2025). Grok 3 Beta — The Age of Reasoning Agents. Accessed at: https://x.ai/news/grok-3. (16 September 2025)

Hadi, M.-U., Qureshi, R., Shah, A., Irfan, M., Zafar, A., Shaikh, M.B., Akhtar, N., Wu, J., Mirjalili, S et al. (2023). *A survey on large language models: Applications, challenges, limitations, and practical usage.* Authorea Preprints. Authorea.

Henriksson, E., Myntti, A., Hellström, S., Erten-Johansson, S., Eskelinen, A., Repo, L. & Laippala, V. (2024). From Discrete to Continuous Classes: A Situational Analysis of Multilingual Web Registers with LLM Annotations. *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pp. 308–318.

Jakubíček, M. & Rundell, M. (2023). The end of lexicography? Can ChatGPT outperform current tools for post-editing lexicography? *Electronic Lexicography in the 21st Century: Invisible Lexicography, Brno, Czechia. eLex 2023*, pp. 518–533.

Karelson, R. (1990). "Eesti kirjakeele seletussõnaraamat" tegija pilgu läbi. *Keel ja Kirjandus*, 1, pp. 24–34.

Kasik, R. (2021). Normikeel ja ühiskeel – eesti keel. *Sirp* 23 July.

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), pp. 7–36.

Klosa-Kückelhaus, A. & Tiberius, C. (2024). The Lexicographic Process Revisited. *International Journal of Lexicography*, 38(1), pp. 1–12.

Koppel, K. & Kallas, J. (2022). Eesti keele ühendkorpuste sari 2013–2021: mahukaim eestikeelsete digitekstide kogu. *Eesti Rakenduslingvistika Ühingu aastaraamat*, 18, pp. 207–228.

Koppel, K., Tavast, A., Langemets, M., & Kallas, J. (2019). Aggregating dictionaries into the language portal *Sõnaveeb*: Issues with and without a solution. *Electronic lexicography in the 21st century. Proceedings of the eLex 2019 conference. 1-3 October 2019, Sintra, Portugal. Lexical Computing*, pp. 1–3.

Koppel, K., Kallas, J., Jürviste, M. & Kaljumäe, H. (2023). *Estonian National Corpus 2023*. Lexical Computing Ltd. / Eesti Keele Instituut.

Langemets, M., Risberg, L. & Algvere, K. (2024). To Dream or Not to Dream About 'Correct' Meanings? Insights into the User Experience Survey. In *XXI EURALEX International Congress. Cavtat, Croatia*, 741−760.

Lew, R. (2024). Dictionaries and lexicography in the AI era. *Humanities and Social Sciences Communications volume 11*. Available at : https://www.nature.com/articles/s41599-024-02889-7. (16 September 2025)

Li, Z., Shi, Y., Liu, Z., Yang, F., Liu, N. & Du, M. (2024). Quantifying multilingual performance of large language models across languages. Available at: https://doi.org/10.48550/arXiv.2404.11553. (16 September 2025)

Marcondes, F.S., Adelino de C. O. S. Gala, Manuel Rodrigues, José João Almeida & Paulo Novais. (2024). Lexicon Annotation with LLM: A Proof of Concept with ChatGPT. *International Conference on Hybrid Artificial Intelligence Systems* (Lecture Notes in Computer Science), pp. 190–200. Available at: https://link.springer.com/chapter/10.1007/978-3-031-74186-9_16. (16 September 2025)

McKean, E. & Fitzgerald, W. (2024). The ROI of AI in lexicography. *Lexicography* 11(1). Available at: https://utppublishing.com/doi/abs/10.1558/lexi.27569. (16 September 2025)

Müller-Spitzer, C. & Koplenig, A. (2014). Online dictionaries: expectations and demands. In C. Müller-Spitzer (ed.). *Using Online Dictionaries. (Lexicographica. Series Maior 145.)* Walter de Gruyter, pp. 143–188.

Nguyen, X.-P., Aljunied, M., Joty, S. & Bing, L. (2024). Democratizing LLMs for Low-Resource Languages by Leveraging their English Dominant Abilities with Linguistically-Diverse Prompts. In L.-W. Ku, A. Martins & V. Srikumar (eds.) *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*. Bangkok, Thailand: Association for Computational Linguistics, pp. 3501–3516.

OpenAI. (2023). GPT-4 Technical Report. Available at: https://openai.com/research/gpt-4. (16 September 2025)

Pajusalu, R. (2009). *Sõna ja tähendus.* Tallinn: Eesti Keele Sihtasutus.

Pullum, G.K. (2023). Why grammars have to be normative – and prescriptivists have to be scientific. In Beal, J., Lukač, M. & Straaijer, R. (2023). *The Routledge Handbook of Linguistic Prescriptivism*. London, New York: Routledge, pp. 3–16.

Raag, R. (2008). *Talurahva keelest riigikeeleks.* Tartu: AS Atlex.

Regular Evaluation Report 2024. Arts and Humanities. Available at: https://etag.ee/wp-content/uploads/2025/05/Eesti-Keele-Instituut.pdf. (16 September 2025)

Risberg, L. (2024). *Sõnatähendused ja sõnaraamat. Kasutuspõhine sisend eesti keelekorraldusele.* (Dissertationes philologiae Estonicae Universitatis Tartuensis 52.) Tartu: Tartu Ülikooli Kirjastus.

Risberg, L., Tuulik, M., Langemets, M., Koppel, K., Vainik, E., Prangel, E. & Aedmaa, E. (2025). Keelekorpus kui leksikograafi abiline kõnekeelsuse tuvastamisel [Using corpus data to support lexicographers in identifying informal language]. *Keel ja Kirjandus* 7, pp. 605—624. Available at: https://doi.org/10.54013/ kk811a3. (16 September 2025)

Rotter, S. & Liu, M. (2023). Interlocutor relation predicts the formality of the conversation: An experiment in American and British English. *Register Aspects of Language in Situation (REALIS)* 2(2), pp. 1–27. Available at: https://doi.org/10.18452/26192. (16 September 2025)

Rundell, M. (2002). Good old-fashioned lexicography: Human judgement and the limits of automation. In M.-H. Corréard (ed.). *Lexicography and Natural Language Processing: A Festschrift in honour of B. T. S. Atkins. Stuttgart: EURALEX*, pp. 138–155.

Rundell, M. (2012). It works in practice but will it work in theory? The uneasy relationship between lexicography and matters theoretical (Hornby Lecture). In Fjeld, R.V. & Torjusen, J.M. (eds.). *Proceedings of the 15th EURALEX Congress. Oslo: University of Oslo*, pp. 47–92.

Rundell, M. (2024). Automating the Creation of Dictionaries: Are We Nearly There? – Humanising *Language Teaching*, 26(1). Available at: https://www.hltmag.co.uk/ feb24/automating-the-creation-of-dictionaries. (16 September 2025)

Sõnaveeb. Language portal, Institute of the Estonian Language. Accessed at: https://sonaveeb.ee. (16 September 2025)

Tao, Y., Viberg, O., Baker, R.S. & Kizilcec, R.F. (2024). Cultural bias and cultural alignment of large language models. In M. Muthukrishna (ed.). *PNAS Nexus*, 3(9).

Trap-Jensen, L. (2002). Descriptive and Normative Aspects of Lexicographic Decision-Making: The Borderline Cases. In *Proceedings of the Tenth EURALEX International Congress. Copenhagen*, pp. 503–509.

Trap-Jensen, L. (2024). The Best of Two Worlds: Exploring the Synergy between Human Expertise and AI in Lexicography. Available at: https://lexicography21.iliauni.edu.ge/wp-content/uploads/2024/06/03_Lars-Trap-Jensen.pdf. (16 September 2025)

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F. and Rodriguez, A. (2023*). Llama: Open and efficient foundation language models*. Available at: https://doi.org/10.48550/arXiv.2302.13971. (16 September 2025)

Tu, N.D.T., Lang, C. & Brunner, A. (2025). LLM fails. Gescheiterte Experimente mit Generativer KI und was wir daraus lernen können. *Workshop am 8. und 9. April 2025, Leibniz-Institut für Deutsche Sprache*. Available at: https://www.ids-mannheim.de/home/lexiktagungen/llm-fails/. (16 September 2025)

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), pp. 147–176. Available at: https://doi.org/10.12697/jeful.2025.16.2.05. (Available in 6–10 October 2025)

Vaik, K. (2024). *Beyond Genres: A Dimensional Text Model for Text Classification.* (Dissertationes linguisticae Universitatis Tartuensis 47.) Tartu: Tartu Ülikooli Kirjastus.

Vare, S. (2001). Üldkeele ja oskuskeele nihestunud suhe. *Keel ja Kirjandus*, 7, pp. 455–472.

Yu, X., Zhang, Z., Niu, F., Hu, X., Xia, X. & Grundy, J. (2024). What Makes a High-Quality Training Dataset for Large Language Models: A Practitioners' Perspective. In *Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering*. New York, NY, USA: Association for Computing Machinery, pp. 656–668.

Choosing Suitable Text Corpora for Identifying Collocations – A Case Study of a Large Reference Dictionary of Contemporary German

Luise Köhler, Gregor Middell, Alexander Geyken

Berlin-Brandenburg Academy of Science and Humanities, Jägerstraße 22/23, 10117 Berlin, Germany

E-mail: luise.koehler@bbaw.de, gregor.middell@bbaw.de, geyken@bbaw.de

Abstract

Collocations are a well-covered research area in lexicography. With the advent of evidence-based lexicography and the availability of large text corpora, computational methods of extracting typical co-occurrences from such corpora and supporting lexicographers in identifying collocations among them became a research focus. Especially the statistical properties of collocations (i.e. application of various association measures) have been evaluated for different languages, collocation types, gold standards and corpora (e.g. Evert et al. 2017; Garcia, García Salido, and Alonso-Ramos 2019). In hindsight though, and despite the undisputed heuristic value of statistical methods for the task at hand, the overall results of such studies do not provide clear conclusions, especially with respect to the practical implications for lexicographic work. Combined, they highlight the dependency of the results on available datasets, investigated collocation types, as well as the underlying corpora in terms of their composition and the affordable preprocessing (Uhrig, Evert, and Proisl 2018). Some results even indicate that for high-quality, dependency-annotated corpora – in contrast to large but scarcely annotated web corpora used in previous studies – raw frequency data can be as indicative for extracting collocations as association measures. Consequently and given recent advances in deep learning, the focus shifted from the evaluation of association measures to the adaptation of increasingly capable statistical language models for the identification and classification of collocations (Espinosa-Anke, Codina-Filbà, and Wanner 2021; Falk et al. 2021; Ljubešić, Logar, and Kosem 2021).

In this study, we examine a more fundamental question that is addressed only in passing by the aforementioned work. This question becomes more important as the focus shifts from the precision of association measures to the recall required when constructing representative datasets for training classifiers: Which type of corpora are actually suitable for extracting collocation candidates and exemplifying their usage? To this end, we compare several corpora of the vast corpus collection of the ‘Digitales Wörterbuch der deutschen Sprache’ (DWDS), that comprises more than 70 billion

tokens of German texts, including reference corpora, web corpora and high-quality print newspapers. In order to study the coverage of collocations by these corpora, we assembled a gold standard from three lexical resources of collocations of contemporary German: the collocations described in DWDS entries, a dictionary of German collocations (Quasthoff 2011), and a dataset from a recent dissertation (Strakatova 2024), yielding in total approximately 350,000 collocations of different syntactic types. We verify the presence of these collocations in various corpora of the DWDS corpus collection. Comparing the coverage of our gold standard datasets by those corpora, we conduct a case study to answer questions such as: a) How good is the coverage of common collocations by carefully selected but small reference corpora? b) Are giga-token web corpora sufficient to cover a broad set of collocations as documented in comprehensive reference dictionaries? c) Do high-quality newspapers surpass web-corpora or can they be replaced by well-curated web corpora?

Keywords: collocations; corpora; evaluation

References

Barbaresi, A. (2021). Trafilatura: A Web Scraping Library and Command-Line Tool for Text Discovery and Extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, pp. 122–131.

Barbaresi, A. (2024). *Simplemma: A Simple Multilingual Lemmatizer for Python* (Version 1.1.2). Berlin, Germany: Berlin-Brandenburg Academy of Science and Humanities. Accessed at: https://doi.org/10.5281/zenodo.14187363. Accessed at: https://github.com/adbar/simplemma.

Bartsch, S. & Evert, S. (2014). Towards a Firthian Notion of Collocation. *Vernetzungsstrategien, Zugriffsstrukturen und automatisch ermittelte Angaben in Internetwörterbüchern*, 2(1), pp. 48–61.

Berlin-Brandenburg Academy of Science and Humanities (n.d.). *DWDS – Digitales Wörterbuch der Deutschen Sprache: Das Wortauskunftssystem zur Deutschen Sprache in Geschichte und Gegenwart*. Accessed at: https://www.dwds.de.

Church, K. & Hanks, P. (1990). Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics,* 16(1), pp. 22–29.

Didakowski, J. & Geyken, A. (2014). From DWDS Corpora to a German Word Profile - Methodological Problems and Solutions. *OPAL - Online publizierte Arbeiten zur Linguistik*, 2(2014), pp. 39–47.

Espinosa-Anke, L., Codina-Filbá, J. & Wanner, L. (2021). Evaluating Language Models for the Retrieval and Categorization of Lexical Collocations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 1406–1417.

Evert, S., Uhrig, P., Bartsch, S. & Proisl, T. (2017). E-VIEW-alation - A Large-Scale Evaluation Study of Association Measures for Collocation Identification. In *Electronic Lexicography in the 21st Century.* In *Proceedings of the eLex 2017 Conference*, pp. 531–549.

Falk, N., Strakatova, Y., Huber, E. & Hinrichs, E. (2021). Automatic Classification of Attributes in German Adjective-Noun Phrases. In *Proceedings of the 14th International Conference on Computational Semantics (IWCS).* Association for Computational Linguistics (ACL), pp. 239–249.

Firth, J.R. (1957). *Papers in Linguistics*, 1934–1951. Oxford University Press.

Garcia, M., García Salido, M. & Alonso-Ramos, M. (2019). A Comparison of Statistical Association Measures for Identifying Dependency-Based Collocations in Various Languages. *Proceedings of the Joint Workshop on Multiword Expressions and Wordnet (MWE-WN 2019).* Association for Computational Linguistics (ACL), pp. 49–59.

Geyken, A. (2007). The DWDS Corpus: A Reference Corpus for the German Language of the 20th Century. In C. Fellbaum (ed.) *Collocations and Idioms: Linguistic, Lexicographic, and Computational Aspects.* Continuum Press London, pp. 23–42.

Goldhahn, D., Eckart, T., Quasthoff, U. et al. (2012). Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pp. 31–43.

Hamp, B. & Feldweg, H. (1997). GermaNet - A Lexical-Semantic Net for German. In *Proceedings of ACL Workshop Automatic Information Extraction and Building of Lexical Semantic Resources.*

Hausmann, F.J. (1984). Wortschatzlernen ist Kollokationslernen. Zum Lehren und Lernen französischer Wortverbindungen. *Praxis des neusprachlichen Unterrichts*, 31(4), pp. 395–406.

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P. & Suchomel, V. (2013). The TenTen Corpus Family. In *7th International Corpus Linguistics Conference CL, Volume 2013*, pp. 125–127.

Kilgarriff, A., Rychlý, P., Smrž, P. & Tugwell, D. (2004). The Sketch Engine. *Proceedings of the 11th EURALEX International Congress*, pp. 105–116.

Köhler, L. & Knaebel, R. (2025). *DWDS Wortprofil* (Version 9.1.0). Zenodo. Accessed at: https://doi.org/10.5281/zenodo.15798452.

Köhler, L., Knaebel, R. & Middell, G. (2025). *German spaCy Models Trained on German UD-HDT and a Collection of German NER Datasets* (Version 2.2.2). Zenodo. Accessed at: https://doi.org/10.5281/zenodo.15797231.

Köhler, L. & Middell, G. (2025). *A German NLP Pipeline for Lexicographic Use Cases* (Version 1.0.1). Zenodo. Accessed at: https://doi.org/10.5281/zenodo.15798301.

Lemnitzer, L., Ermakova, M., Palmes, L., Roll, B., Siebel, K. & Geyken, A. (2025). Kollokationen im DWDS-Wörterbuch und ihr Mehrwert für DaF/DaZ. *Deutsch*

als Fremdsprache, 62(2), pp. 67–80.

Ljubešić, N., Logar, N. & Kosem, I. (2021). Collocation Ranking: Frequency vs Semantics. *Slovenščina 2.0: Empirične, Aplikativne in Interdisciplinarne Raziskave*, 9(2), pp. 41–70.

Quasthoff, U. (2011). *Wörterbuch der Kollokationen im Deutschen*. Walter de Gruyter. Accessed at: https://doi.org/10.1515/9783110225914.

Schmid, H., Fitschen, A. & Heid, U. (2004). SMOR: A German Computational Morphology Covering Derivation, Composition, and Inflection. In *LREC 2004: Fourth International Conference on Language Resources and Evaluation*. European Language Resources Association, pp. 1263–1266.

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford University Press.

Strakatova, Y. (2024). *German Adjective-Noun Co-occurrences with Attributes* [Dataset]. University of Tübingen. Available at: https://doi.org/10.57754/FDAT.76krc-egt63.

Strakatova, Y., Falk, N., Fuhrmann, I., Hinrichs, E. & Rossmann, D. (2020). All that Glitters is not Gold: A Gold Standard of Adjective-Noun Collocations for German. In *Proceedings of the Twelfth Language Resources and Evaluation Conference. European Language Resources Association*, pp. 4368–4378.

Uhrig, P., Evert, S. & Proisl, T. (2018). Collocation Candidate Extraction from Dependency-Annotated Corpora: Exploring Differences across Parsers and Dependency Annotation Schemes. *Lexical Collocation Analysis: Advances and Applications*, pp. 111–140.

A Pipeline for Automated Dictionary Creation with Optional Human Intervention

Thomas Widmann

Danish Language Council, Denmark
E-mail: tw@dsn.dk

Abstract

This paper presents a modular pipeline for automated dictionary creation using large language models (LLMs). It addresses the well-known limitations of prompting systems such as ChatGPT to produce entire entries in a single step – outputs that may read fluently but often lack structural consistency, transparency, originality and verifiability. The proposed system overcomes these weaknesses by decomposing the lexicographic process into a sequence of narrowly constrained, XML-validated stages, each guided by custom-crafted prompts and Document Type Definitions (DTDs). Rather than asking an LLM to "write a dictionary entry," the system treats it as a disciplined assistant performing a defined subtask under strict supervision.

At each stage – ranging from extracting and shortening corpus examples to grouping, defining, translating and formatting – the output is verified against an XML grammar and preserved for audit. This structure enforces reproducibility and allows human intervention at any point, combining the speed and adaptability of machine generation with the oversight and accountability of traditional lexicography. The process is entirely corpus-grounded: every example can be traced to a verifiable source, and every decision in the pipeline is documented. Errors can be corrected where they occur rather than through repeated prompting, and edited intermediate files can be reintegrated seamlessly into the workflow.

Technically, the pipeline is implemented in Python and designed to integrate easily with standard dictionary environments such as IDM's DPS system. It is language-agnostic and domain-independent: prompt files and DTDs can be adapted to any language pair, dictionary type or corpus source. The modular architecture also enables the insertion of new stages – for example, automatic tagging of usage labels, collocations or etymological notes – without altering the underlying structure. The system produces both machine-readable XML output and human-friendly Markdown files for editorial review, ensuring compatibility with established lexicographic and publishing workflows.

Two sample entries for the Danish adjective *nørdet* demonstrate that the pipeline achieves consistent formatting, transparent sourcing and idiomatic translations while avoiding plagiarism and hallucination. Evaluation suggests that each complete run

(typically five stages) produces a usable draft entry at minimal cost and within seconds. The approach therefore provides a sustainable framework for dictionary production, especially for under-resourced languages or specialised terminologies where editorial time and funding are limited.

By embedding formal validation and corpus traceability into every step, the system offers a practical model for responsible integration of LLMs into lexicography. It shifts the human role from mechanical compilation to high-level editorial judgement, enabling lexicographers to supervise, refine and extend AI-generated content with full transparency. Released as open source under the MIT Licence, the pipeline invites adaptation, experimentation and community collaboration.

Keywords: dictionary creation; large language models; lexicographic automation;

XML validation

References

Attardi, G. (2015). WikiExtractor. Accessed at: https://github.com/attardi/wikiextractor.

De Schryver, G.M. & Joffe, D. (2023). The end of lexicography, welcome to the machine. Available at: https://www.youtube.com/watch?v=mEorw0yefAs. Paper presented at the 20th CODH Seminar, National Institute of Informatics, Tokyo, Japan.

Nichols, W. (2023). Invisible lexicographers, AI, & the future of the dictionary. Available at: https://www.youtube.com/watch?v=xYpwftj_QQI. Keynote presented at the eLex conference.

Rundell, M. (2023). Automating the Creation of Dictionaries: Are We Nearly There? In *Proceedings of the 16th International Conference of the Asian Association for Lexicography: Lexicography, Artificial Intelligence, and Dictionary Users*. Seoul: Yonsei University, pp. 1–9.

Tavast, A., Rundell, M., Rychlý, P., Kokol, M. & de Schryver, G.M. (2023). eLex ChatGPT Round Table. Accessed at: https://www.youtube.com/watch?v=dNkksTDYa_s. YouTube video.

Widmann, T. (2023). The Central Word Register of the Danish language. In *Electronic lexicography in the 21st century (eLex 2023): Invisible Lexicography. Proceedings of the eLex 2023 conference*. Brno: Lexical Computing CZ s.r.o., pp. 91–103. Available at: https://elex.link/elex2023/wp-content/uploads/elex2023_proceedings.pdf#page=91.

Widmann, T. (2025a). ai-lexicography: Corpus-based lexicography using the OpenAI. Available at: https://github.com/twidmann/ai-lexicography.

Widmann, T. (2025b). ChatGPT som leksikograf: Ordbogens fremtid? In *20. Møde om Udforskningen af Dansk Sprog*. Aarhus, Denmark: Institut for Kommunikation og

Kultur, Aarhus Universitet, pp. 427–442. Presented October 2024.

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

Jesús Torres del Rey, María García Garmendia

Universidad de Salamanca
E-mail: jtorres@usal.es, mariagarciagarmendia@usal.es

Abstract

While the move to the digital design of lexical resources has, in principle, enhanced the physical and sensory accessibility of dictionaries, a lack of adherence to accessibility standards such as WCAG 2 (Web Content Accessibility Guidelines) (Campbell et all 2023) can introduce significant barriers (NCD 2006; Botelho 2021). These barriers often hinder access to the information and capabilities within those tools or, at the very least, create a user experience that is far from equitable for individuals with disabilities (Lazar et al. 2015: ch. 3, 141; Griffith et al. 2020). However, is formal adherence to standards the only benchmark for actual accessibility to the information, resources and potential knowledge pathways within the e-dictionary?

This study focuses on the accessibility challenges faced by e-dictionary users with visual disabilities. Their exclusion from intellectual or creative tasks frequently stems from ableist perspectives that unjustly assume all-encompassing disabilities for functionally diverse people (Sierra Martínez et al., 2024). However, research has shown that individuals who lack one sense or function often develop remarkable compensatory or divergent abilities (Occelli et al., 2017; Chebat et al., 2020; Sabourin et al., 2022), offering significant potential for professional and intellectual contributions. Yet, they continue to face exclusion in the access to educational and professional contexts due to systemic barriers.

The Diccionario de la Lengua Española (Real Academia de la Lengua 2025), a key reference for the Spanish language, recently underwent a major redesign to achieve state-of-the-art accessibility by aligning with WCAG 2.2 guidelines, particularly as regards programmatic structure and labelling, visual findability and understandability, and use of WAI-ARIA (Accessible Rich Internet Applications) attributes for dynamic content and advanced user interface controls. But does this redesign thoroughly fulfil its accessibility goals?

As proven by users and accessibility experts, and shown in academic literature, a high

score on automated validation tools and strict compliance with guidelines does not necessarily translate into genuine accessibility (Power et al. 2012; Lazar et al. 2015: 153-155). User research is critical in both lexicography (Lew & de Schryver, 2014; Tarp, 2019: 245-246) and accessibility studies (Lazar et al. 2015: ch. 8; Henry et al. 2020). This paper presents an exploratory usability test conducted by a blind user with standard competence in screen reader usage and high academic and professional qualifications, analysed and interpreted by a web accessibility expert. The results identify several areas for improvement in a resource that performs very well in terms of formal accessibility. Examination of actual interaction, however, made us focus on potential problems in usability aspects of the dictionary at the macro and micro structural levels, interaction patterns, and the communication of this information through the assistive technology used, significantly reducing or cancelling their effectiveness (Lew 2012).

Our evaluation methodology combines spontaneous screen reader usability testing, code inspection, and the critical use of automatic validation tools. The results underscore the need for a more user-centred approach to complement existing standards. These findings can contribute not only to advancements in web accessibility standards and practices but also in accessible lexicographic design.

Keywords: web accessibility; usability tests; online dictionaries; user research; screen readers; blind users; visual disabilities; e-dictionary accessibility

References

Botelho, F. H. F. (2021). Accessibility to digital technology: Virtual barriers, real opportunities. *Assistive Technology*, 33(sup1), pp. 27–34. Available at: https://doi.org/10.1080/10400435.2021.1945705.

Campbell, A.; Adams, C.; Montgomery, R. B.; Cooper, M. & Kirkpatrick, A. (eds.) (2023). Web Content Accessibility Guidelines (WCAG) 2.2. W3C Recommendation 05 October 2023.

Chebat D-R.; Schneider, F. C. & Ptito, M (2020) Spatial Competence and Brain Plasticity in Congenital Blindness via Sensory Substitution Devices. *Frontiers in Neuroscience*, 14 815. Available at: doi: 10.3389/fnins.2020.00815.

Griffith, M.; Wentz, B. & Lazar, J. (2020). Measuring the Time Impact of Web Accessibility Barriers on Blind Users: A Pilot Study. In: Langdon, P.; Lazar, J.; Heylighen, A.; Dong, H. (eds.) *Designing for Inclusion*. CWUAAT 2020. Cham: Springer. Available at: https://doi.org/10.1007/978-3-030-43865-4_16.

Henry, S. L.; Abou-Zahra, S. & Arch. A. (2020). Involving Users in Web Projects for Better, Easier Accessibility. Web Accessibility Initiative (WAI). W3C. Available at: https://www.w3.org/WAI/planning/involving-users/.

Lazar J.; Goldstein D. F. & Taylor A. (2015) *Ensuring digital accessibility through*

process and policy. Waltham: Morgan Kaufmann/Elsevier.

Lew, R. (2012). How can we make electronic dictionaries more effective? In: Granger, S & Paquot, M. (eds.) *Electronic Lexicography*. Oxford: Oxford Academic. Available at: https://doi.org/10.1093/acprof:oso/9780199654864.003.0016.

Lew, R. & de Schryver, G. M. (2014). Dictionary users in the digital revolution. International *Journal of Lexicography*, 27(4), pp. 341–359. Available at: https://doi.org/10.1093/ijl/ecu011.

NCD (National Council on Disability). (2006). *Over the Horizon: Potential Impact of Emerging Trends in Information and Communication Technology on Disability Policy and Practice*. Washington, DC: National Council of Disability.

Occelli, V.; Lacey, S.; Stephens, C.; Merabet, L. B & Sathian, K. (2017). Enhanced Verbal Abilities in The Congenitally Blind. *Experimental Brain Research*, 235, pp. 1709–1718. Available at: https://doi.org/10.1007/s00221-017-4931-6.

Power, C.; Freire, A.; Petrie, H. & Swallow, D. (2012). Guidelines are only half of the story: Accessibility problems encountered by blind users on the web. Proceedings of the 2012 CHI Conference on Human Factors in Computing Systems. *ACM Digital Library*, pp. 433–442. Available at: https://doi.org/10.1145/2207676.2207736.

Real Academia española. (2025). Diccionario de la lengua española, 23.ª ed. [online version 23.8]. Accessed at: https://dle.rae.es. Consulted: March, 2025.

Sabourin, C. J.; Merrikhi, Y. & Lomber, S. G. (2022). Do blind people hear better? *Trends in Cognitive Sciences*, 26 (11), pp. 999–1012. Available at: https://doi.org/10.1016/j.tics.2022.08.016.

Sierra Martínez, S.; Fiuza Asorey, M. & Parrilla Latas, A. (2024). Face to face: dialogues around visual impairment. *European Journal of Special Needs Education*, 40 (1), pp. 102–115. Available at: https://doi.org/10.1080/08856257.2024.2331909.

Tarp, S. (2019). Connecting the dots: Tradition and disruption in lexicography. *Lexikos*, 29, pp. 224–249. Available at: https://doi.org/10.5788/29-1-1519.

Mapping Slovene Learner Vocabulary to CEFR Scales with AI-assisted Methods

Mojca Stritar Kučuk

University of Ljubljana, Faculty of Arts, Aškerčeva 2, 1000 Ljubljana
E-mail: mojca.stritarkucuk@ff.uni-lj.si

Abstract

This paper examines how a learner corpus can support lexicographic work by classifying learner vocabulary according to the CEFR scale. Using a corpus-driven methodology, I explore the potential of AI to complement traditional analysis. The study focuses on a selection of texts from the Slovene learner corpus KOST, balanced according to the pragmatically assigned levels of learners' language proficiency: non-Slavic beginners, South Slavic beginners, other Slavic beginners, intermediate and advanced learners. Lemma lists were generated using Sketch Engine and compared with the core vocabulary for Slovene as L2 (up to level B1) and other reference sources. Two advanced language models (ChatGPT and Copilot) were then used to automatically assign CEFR levels to the lemmas. The study compares traditional corpus-derived classifications with AI-generated classifications, evaluates their accuracy and bias, and aims to assess the feasibility of using LLMs in corpus-based CEFR annotation and vocabulary profiling in a lesser-resourced language such as Slovene.

Keywords: CEFR classification; Slovene learner corpus; large language models

(LLMs); vocabulary profiling; second language acquisition

References

Arhar Holdt, Š., Pollak, S., Robnik-Šikonja, M. & Krek, S. (2020). Referenčni seznam pogostih splošnih besed za slovenščino. *Konferenca Jezikovne tehnologije in digitalna humanistika, Ljubljana*, pp. 10–15. Available at: http://nl.ijs.si/jtdh20/pdf/ JT-DH_2020_Arhar-Holdt-et-al_Referencni-seznam-pogostih-splosnih-besed-za-slovenscino.pdf.

Ballier, N., Canu, S., Petitjean, C., Gasso, G., Balhana, C., Alexopolou, T. & Gaillat, T. (2020). Machine learning for learner English. *International Journal of Learner Corpus Research*, 6(1), pp. 72–103.

Berešova, J. (2019). Assigning reference levels to the meanings of words. *EDULEARN19 Proceedings*, pp. 1780–1785. Available at: https://doi.org/ 10.21125/ edulearn.2019.0512.

Cobb, T. & Horst, M. (2015). Learner corpora and lexis. *The Cambridge Handbook of Learner Corpus Research.* Cambridge University Press, pp. 185–206.

Ferbežar, I., Knez, M., Markovič, A., Pirih Svetina, N., Schlamberger Brezar, M., Stabej, M., Tivadar, H. & Zemljarič Miklavčič, J. (2004). *Sporazumevalni prag za slovenščino 2004.* Ljubljana: Center za slovenščino kot drugi/tuji jezik pri Oddelku za slovenistiko Filozofske fakultete Univerze v Ljubljani, Ministrstvo RS za šolstvo, znanost in šport.

Klemen, M., Arhar Holdt, Š., Pollak, S., Kosem, I., Huber, D. & Lutar, M. (2022). Korpus učbenikov za učenje slovenščine kot drugega in tujega jezika. *Na stičišču svetov: slovenščina kot drugi in tuji jezik. Obdobja 41.* Ljubljana: Založba Univerze v Ljubljani, pp. 165–174. Available at: https://doi.org/DOI:10.4312/Obdobja.41.165-174.

Klemen, M., Arhar Holdt, Š., Pollak, S., Kosem, I., Pori, E., Gantar, P. & Knez, M. (2023). Building a CEFR-Labeled Core Vocabulary and Developing a Lexical Resource for Slovenian as a Second and Foreign Language. *Electronic lexicography in the 21st century (eLex 2023): Invisible Lexicography. Proceedings of the eLex 2023 conference.* Brno: Lexical Computing CZ s.r.o., pp. 664–678.

Klinar, M., Pisek, S., Stritar Kučuk, M. & Šter, H. (2022). Poučevanje slovenščine za redno vpisane tuje študente na Univerzi v Ljubljani. *Na stičišču svetov: slovenščina kot drugi in tuji jezik: Obdobja 41.* Ljubljana: Založba Univerze v Ljubljani, pp. 185–194.

Pirih Svetina, N. (2016). *Preživetvena raven za slovenščino: Za potrebe programa opismenjevanje v slovenščini za odrasle govorce drugih jezikov.* Univerza v Mariboru, Filozofska fakulteta, Center za slovenščino kot drugi in tuji jezik. Available at: https://centerslo.si/knjige/ucbeniki-in-prirocniki/prirocniki-in-ucno-gradivo/prezivetvena-raven-za-slovenscino/.

Pitura, J. (2024). Enhancing advanced vocabulary in EFL writing: an AI-assisted intervention for English studies students in Poland. *Journal of China Computer-Assisted Language Learning.* Available at: https://doi.org/10.1515/jccall-2024-0014.

Stritar Kučuk, M. (2024a). KOST 2.0: Predstavitev korpusa in potek označevanja jezikovnih napak. *Zbornik konference Jezikovne tehnologije in digitalna humanistika.* Ljubljana, pp. 589–603. Available at: https://doi.org/10.5281/zenodo.13912515.

Stritar Kučuk, M. (2024b). Investigating the Usage of Machine Translation in L2 Learning and Its Impact on Writing Proficiency. *Lidil* 70. Available at: http://dx.doi.org/10.4000/12lmd.

Stritar Kučuk, M. (2024c). Prvi korpus slovenščine kot tujega jezika KOST 1.0. *Razvoj slovenščine v digitalnem okolju.* Ljubljana: Založba Univerze v Ljubljani, pp. 93–117. Available at: https://doi.org/10.4312/9789612972561.

Svet Evrope (2011). *Skupni evropski jezikovni okvir: Učenje, poučevanje, ocenjevanje.* Ljubljana: Ministrstvo RS za šolstvo in šport, Urad za razvoj šolstva.

Uni, K. (2019). Benefits of Vocabulary of Latin Origin for the Learners of Swedish and

Danish. *The Journal of Social Sciences Research*. Available at: https://doi.org/10.32861/JSSR.52.431.435.

Volodina, E., Pilán, I., Llozhi, L., Degryse, B. & François, T. (2016). SweLLex: Second language learners' productive vocabulary. *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition at SLTC 2016*, *Linköping Electronic Conference Proceedings*, (130), pp. 76–84.

Information seeking behavior of the English learner in the AI era

Anna Dziemianko[1], Mojca M. Hočevar[2]

[1] Adam Mickiewicz University
[2] University of Ljubljana
E-mail: danna@ifa.amu.edu.pl

Abstract

Technology has largely affected the way language learners seek information. Digital formats virtually superseded the paper dictionary (Ptasznik, Wolfer and Lew, 2024), online translators gained much importance (O'Neill, 2019), and web browsers became the first port of call (Kosem et al., 2019). Obviously, generative AI systems imitating human-like communication mark another watershed for online information behavior (De Schryver et al., 2023; Qu and Wu, 2024).

The aim of the study is to investigate English language learners' information seeking behavior on the web in the AI era. The following research questions are posed:

RQ1: Which online tools: search engines/browsers, dictionaries, translators or AI assistants do learners of English access to solve language problems?

RQ2: How often and in what situations do they turn to these tools?

RQ3: How do English learners assess their digital literacy needed to solve language problems using the tools?

RQ4: How do they evaluate the online tools?

RQ5: Which devices are most often used to find linguistic information online?

To answer the research questions, an online questionnaire was designed. So far, it has been conducted among 379 B1/B2+ learners of English in Slovenia, out of whom 161 provided valid answers. Preliminary results indicate that in situations of linguistic deficit, online translators are the first port of call (70%, mainly Google translate, DeepL and Pons), followed by search engines/browsers (60%, mostly Google, less often Safari and Chrome). About 40% of the respondents consult online dictionaries (like the *Cambridge Dictionary*) and AI assistants (ChatGPT, occasionally Deepseek and Grok; RQ1). Online translators and search engines/browsers are typically used once or a few times a week, online dictionaries – once a week or once a month, while AI assistants – every day, once or a few times a week (RQ2). As a rule, all the tools are consulted for both official and unofficial purposes (i.e., to get help with comprehension and

production in daily situations both related and unrelated to university/job). Leisure activities (writing creative texts for pleasure or playing word games) are the least important consultation motives (RQ2). The respondents think a lot of their digital proficiency. Virtually all of them claim that at least half of their last 10 inquiries assisted by any tool were successful (RQ3). Also the tools themselves are highly esteemed. Almost all AI users enjoy their chats, and above 83% of learners like turning to the other tools. However, online dictionaries are considered the most trustworthy (91%), followed by search engines/browsers (68%) and AI assistants (61%). Online translators are trusted the least (53%; RQ4). Interestingly enough, smartphones most often serve to search the web, chat with AI and consult online translators, while online dictionaries are usually accessed from computers (RQ5).

The full paper gives a deeper insight into the tendencies emerging from the collected data, including open-ended questions (e.g., advantages and disadvantages of the investigated tools). The limitations of the study and new avenues of research are also discussed.

Keywords: information seeking behaviour; online dictionary; online translator; web browser/search engine; AI; language learners

References

De Schryver, G.-M. & Joffe, D. (2023). The end of lexicography, welcome to the machine: On how ChatGPT can already take over all of the dictionary maker's tasks. Paper presented at the *20th CODH Seminar. Center for Open Data in the Humanities, Research Organization of Information and Systems, National Institute of Informatics*, Tokyo, 27 February. Accessed at: https://youtu.be/watch?v=mEorw0yefAs.

De Schryver, G.-M. (2023). Generative AI and lexicography: The current state of the art using ChatGPT. *International Journal of Lexicography*, 36(4), pp. 355–387. Available at: https://doi.org/10.1093/ijl/ecad021.

De Schryver, G-M., Rundell, M., Tavast, A., Rychlý, P., Kokol, M. & Krek, S. (2023). Round table on Large Language Models and AI in lexicography. *Panel at the 8th Electronic Lexicography in the 21st Century Conference*. Brno, Czech Republic, 27–29 June. Accessed at: https://www.youtube.com/watch?v=dNkksTDYa_s.

Jakubíček, M. & Rundell, M. (2023). The end of lexicography? Can ChatGPT outperform current tools for post-editing lexicography? In Medveď, M., Měchura, M., Kosem, I., Kallas, J., Tiberius, C. & Jakubíček, M. (eds.), *Proceedings of the eLex 2023 conference*. Brno: Lexical Computing, pp. 518–533.

Kosem, I., Lew, R., Wolfer, S., Müller-Spitzer, C. & Silveira, M. (2019) The image of the monolingual dictionary across Europe: Results of the European survey of dictionary use and culture. *International Journal of Lexicography*, 32(1), pp. 92–

114. Available at: https://doi.org/10.1093/ijl/ecz002.

Levy, M. & Steel, C. (2015) Language learner perspectives on the functionality and use of electronic language dictionaries. *ReCALL*, 27(2), pp. 177–196. Available at: https://doi.org/10.1017/S095834401400038X.

Lew, R. (2023) ChatGPT as a COBUILD lexicographer. *Humanities and Social Sciences Communications*, 10(1), pp. 704. Available at: https://doi.org/10.1057/s41599-023-02119-6.

Nichols, W. (2023) Invisible lexicographers, AI, and the future of the dictionary. Paper presented at the *eLex 2023 Conference: Electronic Lexicography in the 21st Century*. Brno, 27–29 June. Available at: https://www.youtube.com/watch?v=xYpwftj_QQI.

O'Neill, E. (2019) Online translator, dictionary, and search engine use among L2 students. *Computer-Assisted Language Learning Electronic Journal*, 20(1), pp. 154–177.

Ptasznik, B., Wolfer, S. & Lew, R. (2024) A learners' dictionary versus ChatGPT in receptive and productive lexical tasks. *International Journal of Lexicography* (no pagination). Available at: https://doi.org/10.1093/ijl/ecae011.

Qu, K. & Wu, X. (2024) ChatGPT as a CALL tool in language education: A study of hedonic motivation adoption models in English learning environments. *Education and Information Technologies* 29, pp. 19471–19503. Available at: https://doi.org/10.1007/s10639-024-12598-y.

Rees, G. & Lew, R. (2024) The effectiveness of OpenAI GPT-generated definitions versus definitions from an English learners' dictionary in a lexically orientated reading task. *International Journal of Lexicography*, 37(1), pp. 50–74. Available at: https://doi.org/10.1093/ijl/ecad030.

Rundell, M. (2023) Automating the creation of dictionaries: Are we nearly there? In An, Y. (ed.), *Proceedings of the 16th International Conference of the Asian Association for Lexicography*. Seoul: Yonsei University, pp. 1–9.

Steel, C. & Levy, M. (2013) Language students and their technologies: Charting the evolution 2006–2011. *ReCALL*, 25(3), pp. 306–320. Available at: https://doi.org/10.1017/S0958344013000128.

Yang, X. & Yuan, Q. 2022) Six important theories in information behaviour research: A systematic review and future directions. *Information Research, 27*(4): paper 948. Retrieved from http://InformationR.net/ir/27-4/paper948.html (Archived by the Internet Archive at https://bit.ly/3XFayGc). Available at: https://doi.org/10.47989/irpaper948.

Online sources

ChatGPT. Accessed at: https://chatgpt.com.

DeepL. Accessed at: https://www.deepl.com/en/translator.

Deepseek. Accessed at: https://deepseek.ai.

Google translate. Accessed at: https://translate.google.com.

Grok. Accessed at: https://grok.com.

Pons. Accessed at: https://sl.pons.com/prevod; https://en.pons.com/translate.
The Cambridge Dictionary. Accessed at: https://dictionary.cambridge.org.

Making Sense of the Past: AI-Assisted Historical Word Sense Disambiguation and the OED

Elinor Hawkes, Phoebe Nicholson, Will Rogers

Oxford University Press, Great Clarendon Street, Oxford OX2 6DP
E-mail: Elinor.Hawkes@oup.com, Phoebe.Nicholson3@oup.com, Will.Rogers@oup.com

Abstract

This paper presents the *Oxford English Dictionary*'s (*OED*) current exploration into the application of artificial intelligence to historical Word Sense Disambiguation (WSD), a fundamental aspect of *OED*'s core research. Building on a longstanding tradition of technological innovation, the *OED* is investigating how Large Language Models (LLMs) can support the identification and retrieval of illustrative quotations that accurately reflect word sense usage through time – at present one of the most labour-intensive aspects of entry drafting.

The quotation paragraph in *OED* entries provides readers with a curated timeline of usage, illustrating the emergence, evolution, and typical contexts of a word sense. Constructing these paragraphs requires editors to search historical corpora and databases for relevant material, disambiguate search results to isolate the targeted sense, then select quotations that are both representative and informative and meet *OED*'s selection criteria. This task is particularly complex when searching content from earlier time periods, where historical variation in spelling and inflection can further complicate retrieval. Editors currently construct complex iterative search strategies across databases such as *Early English Books Online* (*EEBO*), *Eighteenth Century Collections Online* (*ECCO*), and Google Books, often crafting extensive Boolean queries to find relevant material.

To address these challenges, the *OED* is developing an AI-assisted tool that leverages LLMs to retrieve quotations in specified senses from historical corpora. Rather than relying on manually constructed search strings, the tool allows editors to query the model in natural language, with the LLM returning candidate quotations that match the targeted sense. This approach has the potential to reduce reliance on collocational heuristics, automate the handling of spelling and inflection variants, thus improving the efficiency and accuracy of quotation retrieval.

The paper outlines the technical components of this initiative, including model selection and evaluation, data formatting strategies, prompt engineering strategies, and the quotation retrieval mechanism. Prototype applications are under development to test these components, primarily using *EEBO* as a foundational dataset. Initial testing reveals promising results, though challenges remain, particularly in mitigating LLM

overconfidence and ensuring interpretive caution in ambiguous cases.

In addition to supporting editorial staff, the *OED* is exploring how this tool can benefit subscribers to OED.com. Survey data from academic users indicates strong interest in expanded access to historical quotations, provided the tool is transparent, trustworthy, and well-cited. The paper gives a preview of how the tool might be accessed online, and discusses how the tool might grow from a "Minimum Viable Product" to something more powerful, whilst maintaining the distinction between viewing quotations that have been selected by editors and those that have been automatically retrieved by the tool. The paper concludes by reflecting on the broader potential of AI-assisted WSD in digital humanities research and lexicography, and outlines future directions for development, including expanded corpus coverage and enhanced user functionality.

Keywords: Oxford English Dictionary; word sense disambiguation; AI-powered tooling

Corpus-Based Methods and AI-Assisted Terminography for Contextonym Analysis

Antonio San Martín

University of Quebec in Trois-Rivières, 3351, boulevard des Forges,
Trois-Rivières (Quebec) G8Z 4M3 Canada
E-mail: antonio.san.martin.pizarro@uqtr.ca

Abstract

This paper presents contextonym analysis as a hybrid method combining corpus-based techniques and generative artificial (GenAI) tools to support the writing of precise, context-sensitive terminological definitions. Grounded in the Flexible Terminological Definition Approach, this method is based on the premise that definitions should reflect the most relevant conceptual content activated in specific contexts. Contextonyms (frequent surface co-occurrents within a 50-word window) are extracted in word sketch (WS) form in Sketch Engine and help reveal salient semantic features of a target term without relying on predefined syntactic or semantic relations. The paper outlines strategies for interpreting contextonyms, including filtering concordance lines, consulting WSs, and prompting GenAI tools to assist with interpretation. A typology of contextonyms is proposed, along with a case study illustrating how the method supports the creation of domain-specific definitions. By combining corpus data with AI-assisted interpretation, contextonym analysis offers a robust and user-friendly approach to terminological definition writing.

Keywords: contextonym; terminological definition; word sketch; AI-assisted terminography

1. References

Bowker, L. (2003). Lexical Knowledge Patterns, Semantic Relations, and Language Varieties. *Cataloging & Classification Quarterly*, 37(1–2), 153–171. Available at: https://doi.org/10.1300/J104v37n01_11.

Croft, W. & Cruse, A. (2004). *Cognitive Linguistics*. Cambridge University Press.

Drouin, P. (2010). From a Bilingual Transdisciplinary Scientific Lexicon to Bilingual Transdisciplinary Scientific Collocations. In A. Dykstra & T. Schoonheim (Eds.), *Proceedings of the 14th EURALEX International Congress*. Fryske Akademy, pp. 296–305.

Dubuc, R. (2002). *Manuel pratique de terminologie*. Linguatech.

Evans, V. (2015). A Unified Account of Polysemy Within LCCM Theory. *Lingua*, 157,

100–123. Available at: https://doi.org/10.1016/j.lingua.2014.12.002.

Evans, V. (2019). *Cognitive Linguistics: A Complete Guide*. Edinburgh University Press.

Evert, S. (2009). Corpora and Collocations. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics: An International Handbook, Volume 2*, pp. 1212–1248. Mouton de Gruyter.

Fargas, F. X. (2009). *La definició terminològica* (Termcat, Ed.). Eumo.

Freixa, J. & Fernández-Silva, S. (2017). Terminological Variation and the Unsaturability of Concepts. In P. Drouin, A. Francœur, J. Humbley, & A. Picton (eds.), *Multiple Perspectives on Terminological Variation*, pp. 155–180. John Benjamins. Available at: https://doi.org/10.1075/tlrp.18.07fre.

Gadek, G., Betsholtz, J., Pauchet, A., Brunessaux, S., Malandain, N. & Vercouter, L. (2017). Extracting Contextonyms From Twitter for Stance Detection. *ICAART 2017 - Proceedings of the 9th International Conference on Agents and Artificial Intelligence*, 2, pp. 132–141. Available at: https://doi.org/10.5220/0006190901320141.

Hanks, P. (2020). How Context Determines Meaning. In G. Corpas Pastor & J.-P. Colson (eds.), *Computational Phraseology*, pp. 297–310. John Benjamins. Available at: https://doi.org/10.1075/ivitra.24.15han.

ISO/TC 37/SC 1. (2022). *ISO 704:2022 (Terminology work—Principles and methods)*. ISO.

Jakubíček, M., Měchura, M., Kovář, V. & Rychlý, P. (2018). Practical Post-Editing Lexicography with Lexonomy and Sketch Engine. *XVIII EURALEX International Congress: Lexicography in Global Contexts*. Available at: http://euralex2018.cjvt.si/.

Ji, H., Ploux, S. & Wehrli, E. (2003). Lexical Knowledge Representation with Contexonyms. *Machine Translation Summit IX*, pp. 194–201.

Kecskes, I. (2023). *The Socio-Cognitive Approach to Communication and Pragmatics*. Springer. Available at: https://doi.org/10.1007/978-3-031-30160-5.

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), pp. 7–36. Available at: 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. *Proceedings of the XIII EURALEX International Congress*, pp. 425–432.

Kockaert, H. J. & Steurs, F. (2015). *Handbook of Terminology*. John Benjamins.

Lenci, A. (2008). Distributional semantics in linguistic and cognitive research. *Italian Journal of Linguistics*, 20(1), pp. 1–31.

León-Araúz, P. & San Martín, A. (2018). The EcoLexicon Semantic Sketch Grammar: From Knowledge Patterns to Word Sketches. In I. Kerneman & S. Krek (eds.), *LREC 2018 Workshop "Globalex 2018 – Lexicography & WordNets"*, pp. 94–99. Globalex.

León-Araúz, P., San Martín, A. & Faber, P. (2016). Pattern-Based Word Sketches for

the Extraction of Semantic Relations. In P. Drouin, N. Grabar, T. Hamon, K. Kageura & K. Takeuchi (eds.), *Proceedings of the 5th International Workshop on Computational Terminology*, pp. 73–82.

Meyer, I. (2001). Extracting Knowledge-Rich Contexts for Terminography. A Conceptual and Methodological Framework. In D. Bourigault, C. Christian & M.-C. L'Homme (eds.), *Recent Advances in Computational Terminology*, pp. 279–302. John Benjamins.

Rychlý, P. (2008). A Lexicographer-Friendly Association Score. In P. Sojka & A. Horák (eds.), *Second Workshop on Recent Advances in Slavonic Natural Language Processing, RASLAN 2008*. Masaryk University.

San Martín, A. (2016). *La representación de la variación contextual mediante definiciones terminológicas flexibles* [PhD Thesis, University of Granada]. Available at: https://doi.org/10481/43423.

San Martín, A. (2022a). A Flexible Approach to Terminological Definitions: Representing Thematic Variation. *International Journal of Lexicography*, 35(1), pp. 53–74. Available at: https://doi.org/10.1093/ijl/ecab013.

San Martín, A. (2022b). Contextual Constraints in Terminological Definitions. *Frontiers in Communication*, 7. Available at: https://doi.org/10.3389/fcomm.2022.885283.

San Martín, A. (2024). What Generative Artificial Intelligence Means for Terminological Definitions. In F. Vezzani, G.M. Di Nunzio, B. Sánchez Cárdenas, P. Faber, M. Cabezas García, P. León-Araúz, A. Reimerink & A. San Martín (eds.), *3rd International Conference on Multilingual Digital Terminology Today (MDTT 2024)*. CEUR-WS. Available at: *https://ceur-ws.org/Vol-3703/paper1.pdf*.

San Martín, A. (2025). Optimizing Contextonymic Analysis for Terminological Definition Writing. *Information*, 16(4). Available at: https://doi.org/10.3390/info16040257.

San Martín, A. & Trekker, C. (2021). Adapting Word Sketches for Specialized Knowledge Extraction. In D. Amalia, A.D. Darnis, A. Triatna & D. Khairiah (eds.), *14th International Conference of the Asian Association for Lexicography (ASIALEX)*, pp. 64–87. ASIALEX.

San Martín, A., Trekker, C. & Díaz-Bautista, J. C. (2023). Extracting the Agent-Patient Relation from Corpus With Word Sketches. *Proceedings of the 4th Conference on Language, Data and Knowledge*, pp. 666–675. Available at: https://aclanthology.org/2023.ldk-1.73.pdf.

Seppälä, S. (2015). An Ontological Framework for Modeling the Contents of Definitions. *Terminology*, 21(1), pp. 23–50. Available at: https://doi.org/10.1075/term.21.1.02sep.

Șerban, O., Pauchet, A., Rogozan, A. & Pécuchet, J.-P. (2012). Semantic Propagation on Contextonyms Using SentiWordNet. WACAI 2012, *Workshop Affect, Compagnon Artificiel, Interaction*, pp. 86–94.

Suonuuti, H. (1997). *Guide to Terminology*. Tekniikan Sanastokeskus ry. Available at:

http://www.nordterm.net/wiki/en/index.php/Nordterm_8.

Temmerman, R. (2000). *Towards New Ways of Terminology Description: The Sociocognitive Approach*. John Benjamins.

Vézina, R., Darras, X., Bédard, J. & Lapointe-Giguère, M. (2009). *La rédaction de définitions terminologiques*. Office québécois de la langue française.

Contrasting a new AI-powered dictionary designed for on-screen reading with electronic dictionaries that have evolved from print editions

Ana Frankenberg-Garcia

University of Surrey, UK
E-mail: a.frankenberg-garcia@surrey.ac.uk

Abstract

The use of LLMs in lexicography is a hot topic and indeed the focus of eLex 2025. In the past couple of years, several papers have emerged comparing existing dictionary entries with zero-shot chatbot queries (e.g. Nichols 2023) or with dictionary-like content obtained through the dynamic interaction between experts and chatbots (e.g. Lew 2023, Jakubíček & Rundell 2023). However, studies so far have not appeared to have contrasted well-established dictionaries compiled and edited by lexicographers with new types of dictionaries conceived with AI support.

This paper contrasts a new English dictionary created with the assistance of AI that has been designed for on-screen reading with two prestigious electronic dictionaries that have evolved from print editions. The definitions of 39 lexical items from a text on digital well-being published online in *The Conversation* (Shaleha 2024) were compared in: (a) The Oxford Dictionary of English (ODE), accessed directly from the reading screen by right-clicking on the target item when using an Apple device; (b) the Merriam-Webster Dictionary (MW), accessed via a separate tab from the on-screen reading material; and (c) the new Reverso dictionary, embedded in the reading material through a browser extension.

To focus on vocabulary that readers of English as an additional language might genuinely want to look up, the lexical items included in the analysis were those marked as "off list" in a vocabulary profiling tool (Cobb, n.d.) and in Oxford 3000.

The target items consisted of 13 adjectives, 17 nouns (3 plural) 8 verbs (of which 4 were inflected) and 1 adverbial expression. Part of speech was disambiguated contextually where needed (e.g. *prolonged* was classified as an adjective, not a verb).

To assess the ease of consulting definitions for these items while reading on screen, the three dictionaries were compared according to the following parameters:

1. Coverage (was the target sense provided?)

2. Findability (was the target sense easy to spot?)

3. Readability (how long were the definitions and what vocabulary did they use?)

4. Look-up experience (how straightforward was it to access the dictionary while reading?)

The main differences observed were with regard to the last two of the above. Although Reverso is not immune to known problems of AI in lexicography (Michta & Frankenberg-Garcia, 2025), it outperformed ODE and MW in terms of readability and look-up experience, offering readers short, easy-to-understand definitions that users can consult with minimal disruption while reading electronic texts.

Keywords: LLMs; Vocabulary Assistance; On-screen Reading; Embedded Dictionaries

References

Cobb, T. (n.d.). Compleat Web-VP Classic v.4. Access at: https://www.lextutor.ca/vp/eng/. (25 March 2025)

Jakubíček, M. & Rundell, M. (2023). The end of lexicography? Can ChatGPT outperform current tools for post-editing lexicography? *Proceedings of the eLex 2023 conference*, pp. 518–532. Available at: https://elex.link/elex2023/publications/.

Lew, R. (2023). ChatGPT as a COBUILD lexicographer. Humanities & Social Sciences Communication, 10:704. Available at: https://doi.org/10.1057/s41599-023-02119-6.

Merriam-Webster Dictionary Online. (n.d.). Accessed at: https://www.merriam-webster.com/. (25 March 2025)

Merriam-Webster Dictionary Online. Accessed at: https://www.merriam-webster.com/. (15 July 2025)

Michta, T. & Frankenberg-Garcia, A. (2025). Learners‘ reactions to false polysemy. Paper presented at *eLex 2025.*

Nichols, W. (2023) Invisible lexicographers, AI, & the future of the dictionary. Keynote lecture at eLex 2023. Available at: https://youtu.be/xYpwftj_QQI.

Oxford Dictionary of English App. (n.d.). Access at: https://apps.apple.com/us/app/oxford-dictionary/id978674211. (25 March 2025)

Oxford 3000. Accessed at: https://www.oxfordlearnersdictionaries.com/wordlists/oxford3000-5000. (15 July 2025)

Oxford Dictionary of English. Accessed at: https://apps.apple.com/us/app/oxford-dictionary/id978674211 and https://premium.oxforddictionaries.com/english/. (15 July 2025)

Reverso Dictionary. (n.d). Access at: https://dictionary.reverso.net/english-definition/. (15 September 2025)

Shaleha, R. (2024). Rethinking screen time: A better understanding of what people do on their devices is key to digital well-being. The Conversation, 19 November 2024. Available at: https://theconversation.com/rethinking-screen-time-a-better-understanding-of-what-people-do-on-their-devices-is-key-to-digital-well-being-243644.

Toward a corpus-based multilingual terminology database for Intercultural Communication

María Iglesias Vázquez, Charlotte Venema, Marie Steffens

Utrecht University, Trans 10, 3512 JK Utrecht, The Netherlands
E-mail: mariaiglesiasv00@gmail.com, cacvenema@gmail.com, m.g.steffens@uu.nl

Abstract

This contribution focuses on the methodological aspects of the ICoMuTe project aiming to design a corpus-based multilingual terminology database for Intercultural Communication (ICC). The project seeks to explore how ICC terms relate to each other within six European languages (Dutch, English, German, French, Italian, Spanish), how these terms are connected to their scientific and cultural contexts, and how they can be translated across different languages and cultures while preserving meaning.

The selected approach is corpus-based, using comparable corpora of ICC handbooks and a parallel corpus of texts produced by the European Parliament dealing with key questions related to ICC. Using text recognition and data mining tools (e.g., Sketch Engine), the most frequent ICC terms per language are extracted and analysed in context. To account for the culturally specific aspects of terms while achieving a high degree of cultural neutrality, a semantic model based on tags has been developed for comparing and linking terms across languages in a neutral manner, but natural language corpus-based definitions are also provided that reflect the cultural load of each term.

The main findings suggest that semantic tags are relevant to balance the cultural specificity and neutrality of ICC terms, and that English acts as a reference linguistic and cultural framework for the emergence and development of terms in other languages.

Keywords: intercultural communication; multilingual terminology; corpus-based lexicography; lexical functions; semantic primes

References

Adelstein, A. (2004). *Unidad léxica y valor especializado: estado de la cuestión y observaciones sobre su representación.* [Doctoral research]. Universitat Pompeu Fabra.

Agar, M. (1994). *Language shock. Understanding the culture of conversation.* William Morrow.

Alsina, M.R. (1999). *La Comunicación Intercultural*. Anthropos.

Cabré Castellví, M.T. (1999). *Terminology: Theory, methods and applications*. John Benjamins Publishing Company.

Cambridge Dictionary. (n.d.). *Cambridge Dictionary*. Accessed at: https://dictionary.cambridge.org/. (1 April 2025)

Croft, W. & Cruse, D.A. (2004). *Cognitive linguistics*. Cambridge University Press.

Diki-Kidiri, M. (2022). Cultural Terminology. An introduction to theory and method. In P. Faber & M. L'Homme (Eds.), *Theoretical perspectives on Terminology: Explaining terms, concepts and specialized knowledge*, John Benjamins Publishing Company, pp. 197–226.

Goddard, C. (2004). "Cultural scripts": A new medium for ethnopragmatic instruction. In M. Achard, & S. Niemeier (Eds.), *Cognitive Linguistics, Second Language Acquisition, and Foreign Language Teaching*, Mouton de Gruyter, pp. 145–165.

Goddard, C. & Wierzbicka, A. (1994). *Semantic and lexical universals – Theory and empirical findings*. John Benjamins Publishing Company.

Goddard, C. & Wierzbicka, A. (2002). *Meaning and Universal Grammar – Theory and empirical findings*. John Benjamins Publishing Company.

Goddard, C. & Wierzbicka, A. (2007). Semantic primes and cultural scripts in language learning and intercultural communication. In F. Sharifian, & G.B. Palmer (eds.), *Applied cultural linguistics: Implications for second language learning and intercultural communication*, John Benjamins Publishing Company, pp. 105–124.

Freixa, J. (2006). Causes of denominative variation in terminology. *Terminology*, 12(1), pp. 51–77.

Fuertes-Olivera, P.A. & Tarp. S. (2014). *Theory and practice of specialised online dictionaries: Lexicography versus Terminography*. DeGruyter.

International Organization for Standardization. (2019). *Terminology work and terminology science — Vocabulary* (ISO standard No. 1087:2019).

International Organization for Standardization. (2022). *Terminology work — Principles and methods* (ISO standard No. 704:2022).

Merriam-Webster. (n.d.). *Merriam-Webster Dictionary*. Accessed at: https://www.merriam-webster.com/. (1 April 2025)

Remígio, A.R. (2013). The terminographical process: Phases and dimensions. *Meta*, 58(1), pp. 191–211.

Temmerman, R. (2000). *Towards new ways of terminology description*. John Benjamins Publishing Company.

Vézina, R. (2009). *La rédaction de définitions terminologiques*. Office québécois de la langue française.

Vuorikari, R.H. (2009). *Tags and self-organisation: A metadata ecology for learning resources in a multilingual context* [Doctoral Thesis]. Open Universiteit.

Wierzbicka, A. (1985). *Lexicography and conceptual analysis*. Karoma Publishers.

Žolkovskij, A.K., & Mel'čuk, I.A. (1965). O vozmožnom metode i instrumentax semantičeskogo sinteza (On a possible method and instruments for semantic synthesis). *Naučno-texničeskaja Informacija*, 5, pp. 23–28.

LLM-Assisted Dialect Lexicography: Challenges and Opportunities in Processing Historical Bavarian Dialects

Philipp Stöckle, Daniel Elsner, Wolfgang Koppensteiner, Katharina Korecky-Kröll

Austrian Centre for Digital Humanities, Austrian Academy of Sciences, Bäckerstraße 13, 1010 Vienna, Austria

E-mail: philipp.stoeckle@oeaw.ac.at, daniel.elsner@oeaw.ac.at, wolfgang.koppensteiner@oeaw.ac.at, katharina.korecky-kroell@oeaw.ac.

Abstract

This paper investigates the potential of LLMs in supporting lexicographic work on non-standard linguistic varieties using data from the *Dictionary of Bavarian Dialects in Austria (WBÖ)*. Based on approx. 2.4 million digitized and TEI-encoded dialect paper slips published via the Lexical Information System Austria (LIÖ), we construct a domain-specific corpus and evaluate LLMs in semantic classification and dictionary entry generation. Key preparatory steps include metadata enrichment, glossary and ontology development, and prompt engineering combined with Retrieval-Augmented Generation (RAG) techniques. Preliminary results suggest that LLMs can assist in organizing dialectal material into coherent semantic groupings. However, challenges persist regarding data preprocessing, structural conformity, and selection of representative examples. We discuss methodological implications and outline future directions, including the integration of agent-based systems and fine-tuning approaches tailored to dialect resources. This study contributes to the broader discourse on AI-assisted lexicography, highlighting both the potential and limitations of current LLM technologies in handling underrepresented language varieties.

Keywords: computational lexicography; historical dialect lexicography; large language models; metadata enrichment; semantic classification

References

Baldazzi, T., Bellomarini, L., Ceri, S., Colombo, A., Gentili, A., Sallinger, E. & Atzeni, P. (2023a). Explaining Enterprise Knowledge Graphs with Large Language Models and Ontological Reasoning. Available at: https://doi.org/10.4230/OASIcs.Tannen.1. (18 July 2025)

Baldazzi, T., Bellomarini, L., Ceri, S., Colombo, A., Gentili, A. & Sallinger, E. (2023b). Fine-tuning Large Enterprise Language Models via Ontological Reasoning.

Available at: https://doi.org/10.48550/arXiv.2306.10723. (18 July 2025)

Bowers, J. & Stöckle, P. (2018). TEI and Bavarian dialect resources in Austria: updates from the DBÖ and WBÖ. In A.U. Frank, C. Ivanovic, F. Mambrini, M. Passarotti & C. Sporleder (eds.) *Proceedings of the Second Workshop on Corpus-Based Research in the Humanities (CRH-2), 25–26 January 2018 Vienna, Austria*. Wien: Gerastree Proceedings, pp. 45–54. Available at: https://www.oeaw.ac.at/fileadmin/subsites/academiaecorpora/PDF/CRH2.pdf. (18 July 2025)

Breuer, L.M. & Stöckle, P. (2023). Das WBÖ-online im ‚Lexikalischen Informationssystem Österreich' – Zugriff und Vernetzungsmöglichkeiten, Version 2. In T. Krefeld, S. Lücke & C. Mutter (eds.) *Berichte aus der digitalen Geolinguistik (II): Vernetzung und Nachhaltigkeit (Korpus im Text 9), Version 30*. Accessed at: https://www.kit.gwi.uni-muenchen.de/?p=54448&v=2. (18 July 2025)

Chen, L., Dao, H.-L. & Do-Hurinville, D.-T. (2024). AI empowerment: Where are we in the automation of lexiocography? A metaphraseographic study. In A. Inoue, N. Kawamoto & M. Sumiyoshi (eds.) *ASIALEX 2024 Proceedings*, Sep 2024, Tokyo, Japan, pp. 90–98.

De Schryver, G.-M. (2023). Generative AI and Lexicography: The current State of Art Using ChatGPT. *International Journal of Lexicography* 36(4), pp. 355–387.

Elasticsearch. Accessed at: https://www.elastic.co/elasticsearch. (18 July 2025)

Gupta, S., Ranjan, R. & Narayan Singh, S. (2024). A Comprehensive Survey of Retrieval-Augmented Generation (RAG): Evolution, Current Landscape and Future Directions. Available at: https://doi.org/10.48550/arXiv.2410.12837. (18 July 2025)

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.) *eLex 2023. Electronic lexicography in the 21st century*, Lexical Computing CZ: Brno, pp. 518–532.

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

Li, Z., Shi, Y., Liu, Z., Yang, F., Payani, A., Liu, N. & Du, M. (2024). Language Ranker: A Metric for Quantifying LLM Performance Across High and Low-Resource Languages. Available at: https://doi.org/10.48550/arXiv.2404.11553. (18 July 2025)

LIÖ: *Lexikalisches Informationssystem Österreich* ('Lexical Information System Austria'). Accessed at: https://lioe.dioe.at/. (18 July 2025)

McKean, E. & Fitzgerald, W. (2023). The ROI of AI in Lexicography. In*: AsiaLex 2023. Lexicography, Artificial Intelligence and Dictionary Users*, pp. 18–27.

Meta Llama Model 3 = Meta AI: Introducing Meta Llama 3: The most capable openly available LLM to date. Accessed at: https://ai.meta.com/blog/meta-llama-3/. (18 July 2025)

Llama Team @ Meta: The Llama 3 Herd of Models. Available at:

https://doi.org/10.48550/arXiv.2407.21783. (18 July 2025)

Pan S., Luo L., Wang Y., Chen C., Wang J. & Wu X. (2023). Unifying Large Language Models and Knowledge Graphs: A Roadmap. Available at: https://doi.org/10.48550/arXiv.2306.08302. (18 July 2025)

Phoodai, C. & Rikk, R. (2023). Exploring Capabilities of ChatGPT for Lexicographical Purposes: 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 & S. Krek (eds.) *eLex 2023. Electronic lexicography in the 21st century. Proceedings of the eLex 2023 conference*, Lexical Computing CZ: Brno, pp. 345–375.

Rundell, M. (2023). Automating the creation of dictionaries: are we nearly there? In *AsiaLex 2023. Lexicography, Artificial Intelligence and Dictionary Users*, pp. 9–17.

Stöckle, P. (2021). Wörterbuch der Bairischen Mundarten in Österreich (WBÖ). In A. N. Lenz & P. Stöckle (eds.) *Germanistische Dialektlexikographie zu Beginn des 21. Jahrhunderts*. Stuttgart: Steiner, pp. 11–46. Available at: https://doi.org/10.25162/9783515129206. (18 July 2025)

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E. & Lample, G. (2023). LLaMA: Open and Efficient Foundation Language Models. Available at: https://doi.org/10.48550/arXiv.2302.13971. (18 July 2025)

TUSTEP: *Tuebingen System of Text Processing tools*. Accessed at: https://www.tustep.uni-tuebingen.de/tustep_eng.html. (18 July 2025)

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, N.G., Kaiser, L. & Polosukhin, I. (2017). Attention Is All You Need. Available at: https://doi.org/10.48550/arXiv.1706.03762. (18 July 2025)

Vatsal, S. & Dubey, H. (2024). A Survey of Prompt Engineering Methods in Large Language Models for Different NLP Tasks. Available at: https://doi.org/10.48550/arXiv.2407.12994. (18 July 2025)

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q. & Zhou, D. (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. Available at: https://doi.org/10.48550/arXiv.2201.11903. (18 July 2025)

Wei, X., Wang, S., Zhang, D., Bhatia, P. & Arnold, A. (2021). Knowledge Enhanced Pretrained Language Models: A Compreshensive Survey. Available at: https://doi.org/10.48550/arXiv.2110.08455. (18 July 2025)

Artificial intelligence in English dictionary entries compiled in *Slovar krajšav*

Mojca Kompara Lukančič

University of Maribor
E-mail: mojca.kompara@gmail.com

Abstract

The article describes the use of artificial intelligence in compiling English dictionary entries for a dictionary of abbreviations (*Slovar krajšav*), published in 2025 and financed by the Slovenian Research and Innovation Agency (ARIS). Together with the Slovenian dictionary of abbreviations (*Slovenski slovar krajšav*) published in 2023, the mentioned dictionary adopted a pioneering approach to the compilation of dictionaries in Slovenia; namely, they are the first contemporary dictionaries of abbreviations. The dictionary of abbreviations was compiled in line with an analysis of the characteristics of English dictionary entries for abbreviations and according to the characteristics of the compilation process used for bilingual dictionaries. The dictionary of abbreviations comprises entries in over 20 languages, the most frequent being English, Italian, French etc. In the article, we focus on compiling English dictionary entries and using artificial intelligence as part of that, namely the *Krajšavar* algorithm. The article describes the application of artificial intelligence – the *Krajšavar* algorithm – in the process of compiling the dictionary of abbreviations *Slovar krajšav* and shows the need for a dictionary of abbreviations to be compiled in the Slovenian language. Dictionaries of abbreviations for the Slovenian language are presented in a synchronic and diachronic framework (cf. Kompara Lukančič 2018), namely, two outdated dictionaries *Kratice* (Župančič 1948) and *Rečnik jugoslovenskih skraćenica* (Zidar 1971), and two more recent online dictionary attempts *Slovarček krajšav* (Kompara Lukančič 2006) and *Slovar krajšav* (Kompara Lukančič 2011), and the most recently published Slovenian dictionary of abbreviations *Slovenski slovar krajšav* (Kompara Lukančič 2023). The Slovenian dictionary of abbreviations *Slovenski slovar krajšav* (Kompara Lukančič 2023) led to the compiling of the dictionary of abbreviations *Slovar krajšav*, a collection of 3,500 alphabetically ordered dictionary entries and over 4,200 expansions gathered in a single volume encompassing over 20 foreign languages. In the article, the overall compilation of the dictionary of abbreviations *Slovar krajšav* is presented, and examples of dictionary entries, namely for English abbreviations, are outlined and discussed. As shown by the presented examples, a dictionary entry is composed following the compilation process used in previously published dictionaries *Slovarček krajšav* (Kompara Lukančič 2006), *Slovar krajšav* (Kompara Lukančič 2011) and the Slovenian dictionary of abbreviations *Slovenski slovar krajšav* (Kompara Lukančič 2023), coupled with the characteristics of a range of English dictionaries of abbreviations (Kompara

Lukančič 2009, 2018). The compilation process took almost two decades to complete and included the application of several algorithms, that is, for lemmatisation, language detection, and the automatic recognition of abbreviations. In the final steps of preparation, the dictionary was compiled manually and with the help of AI that permitted abbreviations on a specialised field to be included, as well as relevant abbreviations obtained from a range of texts following the text typology and under the *Krajšavar* algorithm. The dictionary of abbreviations *Slovar krajšav* together with the Slovenian dictionary of abbreviations *Slovenski slovar krajšav* (Kompara Lukančič 2023) therefore represents an important work in the linguistic framework of abbreviations for the Slovenian language.

Keywords: Abbreviations; English; Dictionary; Algorithm; AI; Dictionary of Abbreviations

References

Kompara Lukančič, M. (2025). *Slovar krajšav*. Maribor: Univerza v Mariboru, Univerzitetna založba. V tisku.

Kompara Lukančič, M. (2023). *Slovenski slovar krajšav*. Maribor: Univerza v Mariboru, Univerzitetna založba.

Kompara Lukančič, M. (2018). *Sinhrono-diahroni pregled krajšav v slovenskem prostoru in sestava slovarja krajšav*. Maribor: Univerzitetna založba Univerze.

Kompara Lukančič, M. (2011). *Slovar krajšav*. Kamnik: Amebis, Termania.

Kompara Lukančič, M. (2009). "Prepoznavanje krajšav v besedilih." *Jezikoslovni zapiski,* 15(1–2), pp. 95–112.

Kompara Lukančič, M. (2006). *Slovarček krajšav.* Ljubljana: Inštitut za slovenski jezik Fran Ramovš ZRC SAZU.

Zidar, J. (1971). *Rečnik jugoslovenskih skraćenica*. Beograd: Međunarodna politika.

Župančič, J. (1948). *Kratice: mala izdaja*. Ljubljana: DZS.

Corpus-Based Vocabulary Profiling for Ukrainian: From Lexical Analysis to the PULS Digital Learning Platform

Olena Synchak[1], Vasyl Starko[1], Mariana Burak[1], Mykhaylo Svystun[2]

[1] Ukrainian Catholic University, 2a Kozelnytska Str., Lviv, 79026, Ukraine
[2] independent researcher
E-mail: o_synchak@ucu.edu.ua, v.starko@ucu.edu.ua, mburak@ucu.edu.ua, michael.svystun@gmail.com

Abstract

While CEFR-aligned vocabulary profiles have been developed for many languages (e.g., English, German, and Swedish), Ukrainian as a foreign language (UFL) still lacks an empirically grounded lexical profile. A foundational issue in creating such profiles is combining lexical frequency data with expert knowledge to assign CEFR-level labels. Existing UFL word lists rely primarily on professional expertise rather than systematic data analysis. The development of a Ukrainian vocabulary profile is further complicated by the prevalence of level-straddling textbooks, significant variability of vocabulary across learning materials, and the inherent inflectional complexity of the language. We aim to bridge these gaps by developing a graded word list for UFL learners (CEFR levels A1–C2), using a comprehensive, data-based approach to vocabulary classification.

To this end, we have constructed a one-million-word corpus based on 21 UFL textbooks (A1–C2) using Ukrainian NLP tools and resources, namely the NLP-UK toolkit (github.com/brown-uk/nlp_uk) and the VESUM dictionary (vesum.nlp.net.ua), for automatic tokenization, lemmatization, and morphological tagging. The corpus has yielded a word list of 37,087 lemmas for which both frequency and distributional data (across levels and textbooks) were recorded. This dataset has enabled us to analyze lexical frequency, dispersion, and variability across a representative selection of UFL textbooks.

Another data input was provided by a general-language corpora. We have analyzed lemma frequency data from two Ukrainian corpora (GRAC and BRUK). By integrating frequency data from three corpora with UFL expert analysis, we have assigned CEFR levels to each lexical item and categorized them by part of speech and communicative topic. Crucially, we have applied the significant onset of use approach (Alfter et al., 2016) to address inconsistencies in existing Ukrainian learning materials and achieve a reliable classification.

The paper outlines the methodology for vocabulary extraction, exploration, and

profiling. Expert decision-making follows a two-stage CEFR alignment process to ensure accuracy, consistency, and pedagogically relevant progression. In the external alignment stage, experts independently assign proficiency levels to words. In the internal alignment stage, these assignments are refined by analyzing words within semantic and derivational clusters. This approach proves particularly effective for languages with complex morphology like Ukrainian.

A CEFR-labeled vocabulary profile of 5,891 lexical items, with a target of 10,000 lemmas, developed through in-depth lexical analysis, is published on the PULS platform (puls.peremova.org). It is designed as a digital learning resource with lexical database functionality, allowing word list extraction by CEFR level, thematic group, and part of speech. Currently, A1 and A2 vocabulary items are available, with higher levels in progress. This profile serves as the foundation for the prospective Ukrainian Learner's Dictionary (ULD), which will include detailed lexical entries with part of speech, CEFR label, thematic group, definition at the level of individual senses, corpus-based examples, pronunciation (audio), English equivalents, pictorial illustrations where relevant, and semantic and derivational relations.

The PULS platform fills a critical gap in creating a comprehensive learning system for UFL. Its central component, the Ukrainian Learner's Dictionary, is the first-ever CEFR-labeled corpus-based UFL reference source that will serve the needs of learners, educators, material creators, and proficiency test designers.

Keywords: learner's dictionary; Ukrainian as a foreign language (UFL); vocabulary profile; learning platform; corpus; CEFR

References

Alfter, D., Bizzoni, Y., Agebjórn, A., Volodina, E. & Pilán, I. (2016). From Distributions to Labels: A Lexical Proficiency Analysis Using Learner Corpora. In *Proceedings of the joint workshop on NLP4CALL and NLP for Language Acquisition at SLTC*, 130, pp. 1–7.

Capel, A. (2010). A1-B2 Vocabulary: Insights and Issues Arising from the English Profile Wordlists Project. *English Profile Journal*, 1(1), pp. 1–11.

Pintard, A. & François, T. (2020). Combining Expert Knowledge with Frequency Information to Infer CEFR Levels for Words. In *Language Resources and Evaluation Conference*, pp. 85–92.

Shvedova, M., von Waldenfels, R., Yarygin, S., Rysin, A., Starko, V., Nikolajenko, T. et al. (2016–2025). *GRAC: General Regionally Annotated Corpus of Ukrainian.* Electronic resource: Kyiv, Lviv, Jena. Accessed at: http://uacorpus.org.

Starko, Vasyl, Rysin, Andriy. (2023). Creating a POS Gold Standard Corpus of Modern Ukrainian. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP).* ACL, pp. 91–95. Available at: 10.18653/v1/2023.unlp-1.11.

Volodina, E., Alfter, D. & Lindström Tiedemann, T. (2024). Profiles for Swedish as a

Second Language: Lexis, Grammar, Morphology. In *Proceedings of the Huminfra conference*, pp. 10–19.

From Word of the Year to Word of the Week: Daily-updated Monitor Corpora for 25 Languages

Ondřej Herman[1,2], Miloš Jakubíček[1,2], Jan Kraus[2], Vít Suchomel[1,2]

[1] Lexical Computing, Brno, Czech Republic
[2] Natural Language Processing Centre, Masaryk University, Brno, Czech Re
E-mail: firstname.lastname@sketchengine.eu

Abstract

This paper presents a long-term privately-funded programme focusing on collecting of timestamped monitor corpora in a wide range of (currently 25) languages. These corpora are primarily designed for researching linguistic trends (including neology) and language change over time. They are available through the Sketch Engine platform and vary significantly in size — from 3 million tokens for Irish to over 100 billion tokens for English. The languages currently included are Arabic, Catalan, Chinese, Czech, Danish, Dutch, English, Estonian, French, German, Greek, Hungarian, Italian, Irish, Maltese, Norwegian, Persian, Polish, Portuguese, Russian, Slovak, Slovene, Spanish, Tamil, and Ukrainian; new languages are continuously being added (with Afrikaans, Amharic, Armenian, Azerbaijani, Georgian, Igbo, Indonesian, Oromo, Urdu, Uzbek and Yoruba being the next candidate set of further 10 languages to be added in the coming months).

The corpora are constructed from news articles published on websites worldwide that offer content via newsfeeds (in the form of RSS and Atom formats). Data coverage ranges from as early as 2014 for the oldest corpora to 2023 for the most recently introduced languages. New data is being collected on a daily basis and an update for each trend corpus is published twice a week. The current work builds on the previously published JSI Newsfeed Corpus (Krek & Herman, 2017), which provided news content only until 2022. Since 2021 for English and 2023 for other languages, the data collection process has been carried out independently on the previous work, expanding the number of supported languages and incorporating new data sources. Sketch Engine already contains extra functionalities that are available to corpora with diachronic annotation. Our trend corpora offer analysis on daily, monthly, quarterly or yearly basis, and besides the dedicated Trend function in Sketch Engine (Kilgarriff, 2015) such metadata can be used to refine a lexicographer's analysis in a concordance search, wordlist discovery or collocational behavior of words provided by theWord Sketch feature.

Nearly 30,000 newsfeeds are queried six times a day, yielding up to 180,000 new articles on weekdays and more than 110,000 articles on weekends per day. The publication date is extracted from the information supplied by the feed, ensuring time-stamping as

accurate as possible. The processing pipeline includes several web text cleaning procedures, namely the main text body extraction, removal of near-duplicates, and enriching the data with linguistic annotations, following methodologies similar to those used for the JSI Newsfeed Corpus and the TenTen corpora family (Kilgarriff, 2014).

In addition to corpus construction, the paper details statistics on feed activity – download volumes, the decay rate (how long an existing newsfeed typically lasts to work) – and the most represented websites per language. The paper also showcases examples of functionality offered by Trend corpora that support corpus lexicography and linguistic research, including neologism detection, word sense shift analysis, and timelinebased analysis of trending words and phrases.

Keywords: text corpus; monitor corpus; timestamped texts; trend analysis

References

Davies, M. (2010). The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Literary and linguistic computing*, 25(4), pp. 447–464.

Davies, M. (2017). The new 4.3 billion word NOW corpus, with 4 ‑ 5 million words of data added every day. In *The 9th International Corpus Linguistics Conference, Volume 2017*, pp. 2.

Herman, O. (2025). Automatic Detection of Word Sense Shift from Corpus Data. *Electronic lexicography in the 21st century. Proceedings of the eLex 2025 conference*.

Kilgarriff, A., Baisa, V., Busta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. & Suchomel, V. (2014). The Sketch Engine: Ten Years On. *Lexicography*, 1(1), pp. 7–36.

Kosem, I. (2022). Trendi - a Monitor Corpus of Slovene. In A. Klosa-Kückelhaus, S. Engelberg, C. Möhrs & P. Storjohann (eds.) *Dictionaries and Society. Proceedings of the XX EURALEX International Congress*. Mannheim: IDS-Verlag, pp. 230–239.

Krek, S., Herman, O., Bušta, J., Jakubíček, M. & Novak, B. (2017). JSI Newsfeed corpus. In *The 9th International Corpus Linguistics Conference*. University of Birmingham.

Lin, Y., Michel, J.B., Aiden, E.L., Orwant, J., Brockman, W. & Petrov, S. (2012). Syntactic annotations for the google books ngram corpus. In *Proceedings of the ACL 2012 system demonstrations. Association for Computational Linguistics*, pp. 169–174.

Ó Meachair, M., Bhreathnach, Ú. & Ó Cleircín, G. (2022). Introducing the National Corpus of Irish Project. In T. Fransen, W. Lamb & D. Prys (eds.) *Proceedings of the 4th Celtic Language Technology Workshop within LREC2022*. Marseille, France: European Language Resources Association, pp. 99–103. Available at: https://aclanthology.org/2022.cltw-1.14/.

Pomikálek, J. (2011). *Removing boilerplate and duplicate content from web corpora.*

Ph.D. thesis, Masaryk University.

Suchomel, V., Kraus, J. et al. (2022). Semi-Manual Annotation of Topics and Genres in Web Corpora, The Cheap and Fast Way. In *RASLAN,* pp. 141–148.

Trampus, M. & Novak, B. (2013). Internals of an aggregated web news feed. In *15th Multiconference on Information Society*, pp. 221–224.

The Dictionary of Contemporary Serbian Language (RSSJ): Advanced Automation and Other Challenges

Ranka Stanković, Rada Stijović, Mihailo Škorić, Cvetana Krstev

JERTEH – Language Resources and Technologies Society, Djusina 7, Belgrade, Serbia
E-mail: ranka@jerteh.rs, stijovicr@gmail.com, mihailo@jerteh.rs, cvetana@jerteh.rs

Abstract

This paper introduces the Dictionary of Contemporary Serbian Language (RSSJ), an ongoing large-scale digital lexicographic project designed to serve both human users via web and mobile applications and machines through APIs. Coordinated by the diaspora association "Gathered around the Language" and the Society for Language Resources and Technologies (JeRTeh), RSSJ aims to produce a dictionary of approximately 50,000 frequently used words, reflecting vocabulary used over the past fifty years across diverse functional styles. The headword list is automatically extracted from corpora (SrpKor2013, SrpKor2021), then manually curated and enriched with data from the LeXimirka database. The project implements advanced automation at multiple stages, employing language models and static embeddings (Word2Vec, FastText, Dict2Vec) to identify synonyms, while large language models assisted in generating draft definitions. Additional methods include automated extraction of collocations, syntactic patterns, and exemplary usage via GDEX algorithms, all managed within a DMLex-inspired PostgreSQL data model. The custom web interface enables seamless integration of dictionary editing and corpus querying. Preliminary results demonstrate that automated drafting accelerates to some extent dictionary development, requiring at the same time lexicographers to adopt more dynamic, data-driven workflows and redefine traditional lexicographic practices.

Keywords: dictionary; Serbian language; lexicography; lexicographic database; natural language processing; large language models; word embeddings

References

Abramski, K., Improta, R., Rossetti, G. & Stella, M. (2025). The "LLM World of Words" English free association norms generated by large language models. *Scientific data*, 12(1), pp. 803.

Atkins, B.S. & Rundell, M. (2008). *The Oxford guide to practical lexicography*. Oxford University Press.

Bojanowski, P., Grave, E., Joulin, A. & Mikolov, T. (2017). Enriching word vectors

with subword information. *Transactions of the association for computational linguistics*, 5, pp. 135–146.

De Schryver, G.M. (2024). The Road towards Fine-Tuned LLMs for Lexicography. In *Workshop'Large Language Models and Lexicography'@ EURALEX 2024. ELEXIS Association*, pp. 6–11.

Devlin, J., Chang, M.W., Lee, K. & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, Volume 1 (Long and short papers)*, pp. 4171–4186.

Gantar, A. (2024). Formulating dictionary definitions using artificial intelligence using the example of Slovenian phraseological units [Formulisanje rečničkih definicija pomoću veštačke inteligencije na primeru slovenačkih frazeoloških jedinica]. In S. Marjanović (ed.) *Moderni rečnici u funkciji prosečnoga korisnika: stari problemi, savremeni pravci i novi izazovi, Volume 1 of Leksikografski susreti, chapter 12.* Beograd: Univerzitet u Beogradu, Filološki fakultet, pp. 151–157. 12.

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, pp. 7–36.

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M. & Rychlý, P. (2008). GDEX: Automatically finding good dictionary examples in a corpus. In *Proceedings of the XIII EURALEX international congress, Volume 1.* Universitat Pompeu Fabra Barcelona, pp. 425–432.

Kilgarriff, A., Rychlý, P., Smrž, P. & Tugwell, D. (2004). The Sketch Engine. *Proceedings of the 11th EURALEX International Congress*, pp. 105–116.

Klosa-Kückelhaus, A. & Tiberius, C. (2024). The Lexicographic Process Revisited. *International Journal of Lexicography*, 38(1), pp. 1–12. Available at: https://doi.org/10.1093/ijl/ecae016.

Kosem, I., Krek, S. & Gantar, P. (2020). Defining collocation for Slovenian lexical resources. *Slovenščina 2.0: empirične, aplikativne in interdisciplinarne raziskave*, 8(2), pp. 1–27.

Kosem, I., Krek, S., Gantar, P., Arhar Holdt, Š. & Čibej, J. (2021). Language Monitor: Tracking the Use of Words in Contemporary Slovene. In *Electronic Lexicography in the 21st Century: Post-editing Lexicography. Proceedings of the eLex 2021 Conference*, pp. 514. ELex 2021.

Krek, S. (ed.) (2024). *Large Language Models and Lexicography: Book of Abstracts of the Workshop.* Cavtat, Croatia: Centre for Language Resources and Technologies, University of Ljubljana & ELEXIS Association. Workshop Book of Abstracts.

Krstev, C. (2008). *Processing of Serbian – Automata, Text and Electronic Dictionaries.* Faculty of Philology, Belgrade.

Krstev, C., Pavlović-Lažetić, G. & Obradović, I. (2004). Using Textual and Lexical Resources in Developing Serbian WordNet. *Romanian Journal of Information Science and Technology*, 7(1–2), pp. 147–161.

Lazić, B. & Škorić, M. (2019). From DELA-based dictionary to Leximirka lexical database. *INFOtheca: Journal of Information and Library Science*, 19, pp. 81–98.

Měchura, M. (2024). *Data Structures in Lexicography*. Ph.D. thesis, Ph. D. Thesis. Masaryk University. Brno A multilingual and multifunctional dictionary in the service of language teaching.

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013). Efficient estimation of word representations in vector space. Available at: https://doi.org/10.48550/arXiv.1301.3781.

Navigli, R. & Velardi, P. (2010). Learning word-class lattices for definition and hypernym extraction. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pp. 1318–1327.

Pilehvar, M.T. & Camacho-Collados, J. (2021). Embeddings in Natural Language Processing: Theory and Advances in Vector Representations of Meaning. *Computational Linguistics*, 47(3), pp. 699–701.

Stanković, R., Krstev, C., Lazić, B. & Škorić, M. (2018a). Electronic dictionaries–from file system to lemon based lexical database. In *6th Workshop on Linked Data in Linguistic (LDL-2018), Towards Linguistic Data Science*, pp. 48–56.

Stanković, R., Krstev, C., Stijović, R., Škorić, M. & Gočanin, M. (2021). Towards automatic definition extraction for Serbian. In *EURALEX (Proceedings of the XIX EURALEX Congress of the European Assocition for Lexicography: Lexicography for Inclusion, Volume 2*. Komotini: SynMorPhoSe Lab, Democritus University of Thrace, pp. 695–704.

Stanković, R., Mladenović, M., Obradović, I., Vitas, M. & Krstev, C. (2018b). Resourcebased WordNet augmentation and enrichment. In *Proceedings of the Third International Conference on Computational Linguistics in Bulgaria (CLIB 2018)*, pp. 104–114.

Stanković, R., Radenović, J., Škorić, M. & Putniković, M. (2025). Learning Word Embeddings using Lexical Resources and Corpora. In *Proccedings of 15th International Conference on Information Society and Technology – ICIST 2025*.

Stijović, R., Krstev, C. & Stanković, R. (2021). Automatska ekstrakcija definicija – doprinos ubrzanju izrade rečnika / Automatic Definition Extraction – Accelerating Dictionary Development. In *Leksikologija i leksikografija u svetlu aktuelnih problema / Lexicology and Lexicography in the Light of Current Issues*. Beograd: Institut za srpski jezik SANU, pp. 113–137.

Stijović, R., Stanković, R. & Škorić, M. (2025). Dictionary of Modern Serbian Language: RSSJ. In *Book of Abstracts of the International Conference South Slavic Languages in the Digital Environment – JuDig*, pp. 32.

Tiberius, C., Kallas, J., Koeva, S., Langemets, M. & Kosem, I. (2024). A Lexicographic Practice Map of Europe. *International Journal of Lexicography*, 37(1), pp. 1–28.

Tufis, D., Cristea, D. & Stamou, S. (2004). BalkaNet: Aims, methods, results and perspectives. A general overview. *Romanian Journal of Information Science and Technology*, 7(1–2), pp. 9–43.

Vitas, D. & Krstev, C. (2012). Processing of corpora of serbian using electronic

dictionaries. *Prace Filologiczne*, LXIII, pp. 279–292.

Vitas, D. & Krstev, C. (2015). Nacrt za informatizovani reˇcnik srpskog jezika. [Blueprint for the computerized dictionary of the Serbian language (in Cyrillic).]. *Naučni sastanak slavista u Vukove dane – Srpski jezik i njegovi resursi: teorija, opis i primene*, 44(3), pp. 105–116.

Vitas, D., Stanković, R. & Krstev, C. (2025). The Many Faces of SrpKor. In *Proceedings of the International Conference South Slavic Languages in the Digital Environment - JuDig*, pp. 1–28.

Learners' reactions to false polysemy

Tomasz Michta[1], Ana Frankenberg-Garcia[2]

[1] University of Białystok

[2] University of Surrey

E-mail: t.michta@gmail.com, ana.frankenberg@gmail.com

Abstract

Studies comparing dictionary entries generated with AI with those of well-established dictionaries edited by lexicographers show that LLMs tend to perform better in some tasks (e.g. writing definitions) than in others (e.g. word-sense disambiguation (e.g. Nichols 2023, Lew 2023, Jakubíček & Rundell 2023, Rees & Lew 2024). One of the problems resulting from the latter is that of "false polysemy" (Jakubíček & Rundell 2023: 525), where the differences between senses listed under a headword are unclear.

Admittedly, the separation of meanings in dictionaries is artificially drawn by lexicographers, and there are often mismatches in sense distinctions across dictionaries. Yet it is still possible for experts to evaluate whether meaning boundaries are sufficiently clear-cut. What is less known is how learners react to false polysemy. Granted that people rarely read dictionary entries in full (Tono 1984, Nuccorini 1994, Bogaards 1998, Dziemianko 2016), and have been reported to stop reading once they find the information they need (Lew, Grzelak & Leszowicz 2013), we wanted to explore whether false polysemy disrupts the consultation process.

This study analysed how 98 L2-English undergraduate students reacted to false polysemy. They took an online quiz consisting of 20 unknown lexical items presented in the context of sentences selected from corpora, some of which were shortened or slightly edited to remove contextual clues. For each vocabulary test item, the participants were given two definitions copied from Reverso, a new English dictionary developed with the assistance of LLMs. Example sentences and sense indicators that could give additional cues about meaning were deliberately omitted. For half of the test items, the pair of definitions provided were indisputably different. For the other half, the definitions were not clearly distinct according to two independent experts (i.e., they were exemplars of AI-generated false polysemy). The test items were shown to the participants in a random order, and each time they were asked to select which of the two definitions (also randomly ordered) was a better fit. They were then asked to judge on a Likert scale how confident they were that they had selected the correct sense. We also recorded the time spent on each test item, the order of the definition selected (first or second), and whether it was correct (when senses were distinct). A sample of the participants was then interviewed to gain further insights into their reactions.

Preliminary results indicate that the participants had little difficulty selecting the

correct sense in the true polysemy condition. However, when faced with false polysemy, their confidence dropped and they took longer to decide. Both effects were statistically significant. Our findings suggest that false polysemy can be detrimental to the user experience, and underscore the need for AI-powered systems that acknowledge and address the problem proactively, as recognized by the developers of Reverso, where human expertise, editorial guidelines and built-in feedback loops are key. That said, future user studies on false polysemy require naturalistic observations, as dictionary users may react differently when not explicitly asked to pick one out of two controlled definitions.

Keywords: LLMs in lexicography; False Polysemy; User Studies

References

Bogaards, P. (1998). 'Scanning Long Entries in Learners' Dictionaries' In Fontenelle Thierry, Hiligsmann Philippe, Michiels Archibald, Moulin André, Theissen Siegfried (eds), *EURALEX'98 Proceedings*. Liège: University of Liège, pp. 555–563.

Dziemianko, A. (2016). Dictionary Entries and Bathtubs: Does It Make Sense?, *International Journal of Lexicography*, 30(3), pp. 263–284. Available at: https://doi.org/10.1093/ijl/ecw010.

Jakubíček, M. & Rundell, M. (2023). The end of lexicography? Can ChatGPT outperform current tools for post-editing lexicography*? Proceedings of the eLex 2023 conference*. Brno: Lexical Computing CZ s.r.o., pp. 518–532. Available at: https://elex.link/elex2023/publications/.

Lew, R. (2023). ChatGPT as a COBUILD lexicographer. *Humanities & Social Sciences Communication*, 10, pp. 704. Available at: https://doi.org/10.1057/s41599-023-02119-6.

Lew, R., Grzelak, M. & Leszowicz, M. (2013). How dictionary users choose senses in bilingual dictionary entries: An eye-tracking study. *Lexikos,* 23, pp. 228–254.

Nichols, W. (2023). Invisible lexicographers, AI, & the future of the dictionary. Keynote lecture at *eLex 2023*. Available at: https://youtu.be/xYpwftj_QQI.

Nuccorini, S. (1994). On Dictionary Misuse. In Willy Martin, Meijs Willem (eds), EURALEX'94: *Proceedings of 6th EURALEX*. Amsterdam: Vrije Universiteit, pp. 586–597.

Rees, G. P. & Lew, R. (2024). The Effectiveness of OpenAI GPT-Generated Definitions Versus Definitions from an English Learners' Dictionary in a Lexically Orientated Reading Task. *International Journal of Lexicography*, 37(1), pp. 50–74. Available at: https://doi.org/10.1093/ijl/ecad030.

Reverso Define. (n.d.). Accessed at: https://dictionary.reverso.net. (31 March 2025)

Tono, Y. (1984). *On the Dictionary User's Reference Skills*. B.Ed. Thesis. Tokyo: Tokyo Gakugei University.

So Close but Still Far: Case Study on Application of LLMs in Idioms Identification, Definition, and Generation of Illustrative Examples

Aleksandra Marković[1], Ranka Stanković[2]

[1] Institute for the Serbian Language SASA, Belgrade, Serbia
[2] University of Belgrade, Faculty of Mining and Geology, Belgrade, Serbia
E-mail: aleksandra.markovic@isj.sanu.ac.rs, ranka.stankovic@rgf.bg.ac.rs

Abstract

Automation has revolutionised lexicography, introducing the 'post-editing lexicography' model, where the role of the lexicographer involves refining automatically generated dictionary drafts. Since the launch of ChatGPT in November 2022, numerous papers have explored the potential applications of LLMs in dictionary production. The rapid evolution of LLMs necessitates a re-evaluation of conclusions drawn approximately two years prior regarding their application in automating dictionary entry creation, particularly in light of the advanced capabilities demonstrated by contemporary models.

We will illustrate an experiment conducted on a dataset of 400 (397) MWEs with idiomatic meaning, aiming to evaluate the usefulness of LLMs in Serbian descriptive lexicography tasks (idiom generation, word-sense disambiguation of MWEs, definition writing, and generation of illustrative examples). We requested two types of illustrative examples: those in which a MWE has an idiomatic meaning, and examples with that meaning paraphrased literally (without the idiom). We will highlight the challenges and issues encountered with several models (ChatGPT-4o and 4.1, Gemini-2.5-Flash and 2.5-Pro) and discuss the differences in their performance based on given LLM prompts using direct chat and APIs access via Python scripts.

Keywords: descriptive monolingual lexicography; LLMs; word sense disambiguation; generating idioms, definitions and illustrative examples; Serbian language

References

Beliga, S. & Filipović Petrović, I. (2024). Large Language Models Supporting Lexicography: Conceptual Organization of Croatian Idioms. In *Proceedings of the Conference on Language Technologies and Digital Humanities*, pp. 23–46.

De Schryver, G.-M. (2024). The Road towards Fine-Tuned LLMs for Lexicography.

Large Language Models and Lexicography, pp. 6.

De Schryver, G.-M. & Joffe, D. (2023). The end of lexicography, welcome to the machine: on how ChatGPT can already take over all of the dictionary maker's tasks. Available at: https://www.youtube.com/watch?v=mEorw0yefAs.

Filipović-Petrović, I. & Beliga, S. (2024). Lexicographic Treatment of Idioms and Large Language Models: What Will Rise to the Surface? In *Book of Abstracts of the Workshop Large Language Models and Lexicography*. Ljubljana: Centre for language resources and technologies, University of Ljubljana, pp. 12–16.

Gantar, A. (2024). Formulisanje rečničkih definicija pomoću veštačke inteligencije na primeru slovenačkih frazeoloških jedinica. In S. Marjanović (ed.) *Moderni rečnici u funkciji prosečnoga korisnika: stari problemi, savremeni pravci i novi izazovi*. University of Belgrade, Faculty of Philology, pp. 151–157.

Gortan Premk, D. et al. (1959–2023). *Rečnik srpskohrvatskog književnog i narodnog jezika SANU, I–XXII [The Dictionary of the Serbo-Croatian Standard and Vernacular Language]*. Beograd: Institut za srpski jezik SANU i SANU. 22 volumes.

Jakubiček, M. & Rundell, M. (2023). The end of lexicography? Can ChatGPT outperform current tools for post-editing lexicography. In *Electronic lexicography in the 21st century (eLex 2023): Invisible Lexicography. Proceedings of the eLex 2023 conference*. pp. 518–533.

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M. & Rychlý, P. (2008). GDEX: Automatically finding good dictionary examples in a corpus. In *Proceedings of the XIII EURALEX international congress, Volume 1*. Universitat Pompeu Fabra Barcelona, pp. 425–432.

Kovačević, Ž. (2002). *Srpsko-engleski frazeološki rečnik*. " Filip Višnjić".

Krstev, C. & Stanković, R. (2023). *European Language Equality: A Strategic Agenda for Digital Language Equality*, chapter Language Report Serbian. Cham: Springer International Publishing, pp. 203–206. Available at: https://doi.org/10.1007/978-3-031-28819-7_32.

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

Marković, A. & Stanković, R. (2025). Primena velikih jezičkih modela u srpskoj opisnoj leksikografiji – studija slučaja. In M. Dinić Marinković & B. Kovačević (eds.) *Applied Linguistics Today – Modern Approaches to Old and New Challenges*. University of Belgrade – Faculty of Philology. Accepted.

Otašević, Đ. (2012). *Frazeološki rečnik srpskog jezika*. Prometej.

Phoodai, C., Rikk, R., Medveď, M., Měchura, M., Kosem, I., Kallas, J., Tiberius, C. & Jakubíček, M. (2023). Exploring the Capabilities of ChatGPT for Lexicographical Purposes: A Comparison with Oxford Advanced Learner's Dictionary within the Microstructural Framework. In *Electronic lexicography in the 21st century (eLex 2023): Invisible Lexicography. Proceedings of the eLex2023 conference*. pp. 345–375.

Rundell, M. (2024). Automating the creation of dictionaries: are we nearly there?

Humanising Language Teaching, 26(1).

Stanković, R., Šandrih, B., Stijović, R., Krstev, C., Vitas, D. & Marković, A. (2019). SASA dictionary as the gold standard for good dictionary examples for Serbian. *Electronic lexicography in the 21st century: Smart lexicography*, pp. 248–269.

Stevanović, M. et al. (ed.) (1982). *Matica srpska*; *Matica hrvatska*, fototipsko izd. edition. Knjige 1, 2, 3 - zajedničko izd. Matice srpske i Matice hrvatske; knj. 4, 5, 6 – izdanje Matice srpske.

Vision-Enabled Language Models in Lexicographical Digitisation: A Case Study of Anton Thor Helle's 1732 Dictionary

Madis Jürviste[1,2], Tiina Paet[1]

[1] Institute of the Estonian Language
[2] University of Tartu
E-mail: madis.jyrviste@eki.ee, tiina.paet@eki.ee

Abstract

Traditionally, historical texts' optical character recognition (OCR) has primarily been conducted using specialised software such as Transkribus, eScriptorium, Kraken, and similar tools. To achieve accurate character recognition, these systems require extensive pre-training and the creation of a refined "ground truth" dataset. The comprehensiveness of model pre-training directly correlates with the precision of results. Large language models (LLMs) promise a potential breakthrough in this domain, offering high-quality output without pre-training through their "zero-shot" capabilities.

Within the framework of a dedicated research programme, "Application of Large Language Models in Lexicography: New Opportunities and Challenges", we have conducted experiments employing untrained language models for the optical character recognition and data structuring of the dictionary section of Anton Thor Helle's 1732 grammar. The recent introduction of vision-capable language models proved decisive, enabling significantly more efficient processing of scanned documents than previously possible.

Preliminary tests demonstrated that Anthropic's Claude 3.5 Sonnet model could generate a structured table from a scanned dictionary file containing Gothic script (Fraktur) based on a simple prompt, recognising the text and appropriately categorising headword entries into relevant columns. Our comparative analysis of various generative language models (Anthropic's Claude, OpenAI's GPT models, Google's Gemini 2.0, and Mistral) revealed that Claude significantly outperforms other models in processing 17th and 18th-century Estonian texts printed in Gothic typeface. Following our preliminary experiments, Anthropic released Claude Sonnet version 3.7, with which we conducted a more comprehensive test to digitise Helle's entire dictionary.

Our presentation examines how effectively the language model transforms a scanned dictionary into a structured, editable document. We assess the accuracy of character recognition for Estonian headwords, German equivalents, and expressions at both character and word levels (CER and WER, respectively) and the precision of data

structuring. Additionally, we explore the most common errors made by the model, factors influencing recognition accuracy, and challenges in adherence to provided prompt instructions.

Claude achieved the highest recognition accuracy with German translation equivalents, as it possesses substantially more training data for German than for Estonian. With both Estonian headwords and German equivalents, Claude frequently modernised word forms. In some instances, the LLM produced "hallucinations" that appeared plausible but bore no relation to the original text. In essence, the LLM tidied the image according to its own understanding — a tendency also observed in experiments with Stahl, Gutslaff, and Göseken (Author 1, Author 2, Author 3, 2025).

The primary advantage of our approach over conventional OCR methods lies in the significant time savings, considering both character recognition and automatic post-structuring capabilities. Whilst the classical method requires extensive ground truth creation and sometimes manual text segmentation, the language model-based approach delivers excellent results with substantially less preparation. Even paid language models such as Claude 3.7 Sonnet prove highly cost-effective.

LLM-based character recognition (and, when necessary, automatic post-structuring) can be applied to digitising other historical texts where prevalent methods would be impractical due to time constraints. This opens new prospects for digitising historical textual heritage and creates prerequisites for more extensive research of old textual sources.

Keywords: historical lexicography; LLMs; OCR

References

Helle, A.T. (1732). *Kurtzgefaszte Anweisung Zur Ehstnischen Sprache*. Halle: Stephan Orban. Available at: http://www.digar.ee/id/nlib-digar:100071.

Author 1; Author 2. (2025). Veebirakenduse Anthonius repositoorium. [Repository of the web application Anthonius.] Available at: https://github.com/joonatanjak/anthonius. (24 April 2025).

Author 1; Author 2; Author 3. (2025) (in press). Eesti vanade sõnakujude tuvastamise võimalustest suurte keelemudelite abil [Identifying Old Estonian Word Forms Using Large Language Models]. Eesti Rakenduslingvistika Ühingu aastaraamat, 21.

How Effective is AI as a Language Consultant?

Urška Vranjek Ošlak

ZRC SAZU, Fran Ramovš Institute of the Slovenian Language
E-mail: urska.vranjek@zrc-sazu.si

Abstract

This paper explores the applicability of generative artificial intelligence in the field of language consulting, focusing on ChatGPT-4 and the Slovenian language. The analysis is based on an experiment involving 30 real user questions submitted to the Language Consulting Service (LCS) of the Fran Ramovš Institute of the Slovenian Language. The questions cover a range of linguistic categories and were submitted to ChatGPT under controlled conditions. The responses were then compared with expert-produced answers and evaluated in terms of factual accuracy, stylistic appropriateness, terminological correctness, and overall usefulness. The results show that while ChatGPT performs well in terms of clarity, tone, and structure, its output often contains inaccuracies and occasionally misleading information. At this stage, ChatGPT is not suitable as a stand-alone tool for end-users. However, it could serve as a helpful draft generator for human language consultants. The study also outlines ways to improve AI output, including better prompts and access to relevant databases. Although some fundamental limitations of AI remain, its controlled use in language consulting may offer practical support, especially in cases involving repetitive or less complex queries.

Keywords: generative AI; language consulting; applied linguistics; Slovenian language

References

Belda-Medina, J. & Calvo-Ferrer, J.R. (2022). Using chatbots as AI conversational partners in language learning. *Applied Sciences*, 12(17), 8427.

Benko, V., Kříhová, Z., Lehečka, B. & Vystrčilová, D. (2024). Persian to Czech Dictionary – A Traditional Dictionary in the Era of AI? In K.Š. Despot, A. Ostroški Anić & I. Brač (eds.) *Lexicography and Semantics: Proceedings of the 21st EURALEX International Congress*, pp. 565–581.

Carbajal-Carrera, B. (2024). AIsplaining: Generative AI explains linguistic identities to me. *Australian Review of Applied Linguistics*, 47(3), pp. 340–365.

De Schryver, G.-M. (2023). Generative AI and lexicography: The current state of the art using ChatGPT. *International Journal of Lexicography*, 36(4), pp. 355–387.

De Schryver, G.-M., Chishman, R. & da Silva, B. (2019). An overview of digital lexicography and directions for its future: An interview with Gilles-Maurice de Schryver. *Calidoscópio*, 17(3), pp. 659–683.

Dobrovoljc, H., & Vranjek Ošlak, U. (2021). Codification within reach: Three clickable layers of information surrounding the new Slovenian normative guide. In I. Kosem, M. Cukr, M. Jakubíček, J. Kallas, S. Krek & C. Tiberius (eds.) *Electronic lexicography in the 21st century: Post-editing lexicography. Proceedings of the eLex 2021 conference*, pp. 637–652.

Dobrovoljc, H., Lengar Verovnik, T., Vranjek Ošlak, U., Michelizza, M., Weiss, P. & Gliha Komac, N. (2020). *Kje pa vas jezik žuli?* Ljubljana: Založba ZRC, ZRC SAZU.

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 (eLex 2023): Invisible Lexicography. Proceedings of the eLex 2023 conference*, pp. 518–533.

Kosem, I., Gantar, P., Arhar Holdt, Š., Gapsa, M., Zgaga, K. & Krek, S. (2024). AI in lexicography at the University of Ljubljana. In S. Krek (ed.) *Book of abstracts of the workshop "Large language models and lexicography"*. Cavtat, Croatia, p. 29.

Krvina, D. & Vranjek Ošlak, U. (2025). *Nadpomenka za versko stavbo*. Jezikovna svetovalnica. Accessed at: https://svetovalnica.zrc-sazu.si/topic/7427/nadpomenka-za-versko-stavbo. (20 July 2025)

Lew, R. (2015). Dictionaries and their users. In H. Jackson (ed.) *The Bloomsbury Companion to Lexicography*. London: Bloomsbury Academic, pp. 244–265.

Ludányi, Z. (2020). Language consulting: A brief European overview. *Eruditio – Educatio*, 15(3), pp. 25–47.

Mžourková, H. (2024). Czech and Slovenian Language Consulting: What Do They Tell Us about Language and Language Users? In S. Štumberger (ed.) *Predpis in norma v jeziku*. Obdobja 43. Ljubljana: Založba Univerze v Ljubljani, pp. 225–231.

Palma, G. (2025). *All ChatGPT Models Explained: Which One Should You Use?* Accessed at: https://justainews.com/blog/openai-chatgpt-models-explained/. (23 May 2025)

Petrič, T., Arhar Holdt, Š. & Robnik-Šikonja, M. (2024). Pomembnost realistične evalvacije: Primer popravkov sklona in števila v slovenščini z velikim jezikovnim modelom. *Slovenščina 2.0*, 12(1), pp. 106–130.

Ptasznik, B. & Lew, R. (2025). Dictionaries versus AI Tools through the Eyes of English Majors. *International Journal of Lexicography*, XX, pp. 1–19.

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.

Rundell, M. (2023). Automating the Creation of Dictionaries: Are We Nearly There? In *Proceedings of the 16th International Conference of the Asian Association for Lexicography (ASIALEX 2023)*, Seoul, Korea, 22–24 June 2023, pp. 9–17.

Saravia, E. (2025). *Prompt engineering guide*. Accessed at: https://www.promptingguide.ai/. (20 June 2025)

Vajjala, S. (2024). Generative Artificial Intelligence and Applied Linguistics. *JALT Journal*, 46(1), pp. 55–76.

Vranjek Ošlak, U. (2023). Language counselling: Bridging the gap between codification and language use. *Eesti ja Soome-Ugri Keeleteaduse Ajakiri / Journal of Estonian and Finno-Ugric Linguistics*, 14(1), pp. 149–173.

Vranjek Ošlak, U. (2024). Jezikovno svetovanje v obdobju prenove pravopisnega priročnika. In S. Štumberger (ed.) *Predpis in norma v jeziku*. Obdobja 43. Ljubljana: Založba Univerze v Ljubljani, pp. 399–405.

Žaucer, R. & Marušič, F. (2009). Jezikovno svetovanje, praksa in ideali. In M. Stabej (ed.) *Infrastruktura slovenščine in slovenistike*. Obdobja 28. Ljubljana: Založba Univerze v Ljubljani, pp. 449–456.

Lexical-Semantic Resources as a Culture-Aware Basis for Benchmarking and Evaluation of LLMs

Nathalie Norman[1], Sanni Nimb[2], Sussi Olsen[1],

Nina Schneidermann[1], Bolette S. Pedersen[1]

[1] University of Copenhagen, Njalsgade 80, DK-2300 S
[2] The Society for Language and Literature, Christians Brygge 1, DK-1219 K
E-mail: naha@hum.ku.dk, sn@dsl.dk, saolsen@hum.ku.dk, ninasc@hum.ku.dk, bspeder-sen@hum.ku.dk

Abstract

Large Language Models (LLMs) tend to expose severe language and cultural biases when working in medium- and low-resourced languages. In this paper, we present our work on Danish benchmarking and evaluation of LLMs to more precisely diagnose and potentially remedy such bias. To this aim, we apply available lexical-semantic resources to compile a set of Natural Language Understanding (NLU) tasks in Danish that reflect the breadth and nuances of the Danish vocabulary, thereby capturing also implicit traits of Danish values and culture. Currently the benchmark comprises nine NLU tasks, including tasks such as disambiguating words in context, determining semantic outliers, inferencing and interpretation tasks based on semantic relations, as well as selecting the correct explanation of culture-related metaphorical idioms. The large-scale benchmark (currently approx. 8,000 data instances) is supplemented by a selection of a much smaller dataset prepared for human evaluation of LLM-generated explanations, thereby enabling a more careful study of the language generation and interpretation abilities of the models from a lexical-semantic perspective.

Keywords: benchmarks; Large Language Models; evaluation; lexical resources

References

Berdicevskis, A., Bouma, G., Kurtz, R., Morger, F., Öhman, J., Adesam, Y., Borin, L., Dannélls, D., Forsberg, M., Isbister, T., Lindahl, A., Malmsten, M., Rekathati, F., Sahlgren, M., Volodina, E., Börjeson, L., Hengchen, S. & Tahmasebi, N. (2023). Superlim: A Swedish language understanding evaluation benchmark. In H. Bouamor, J. Pino & K. Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 8137–8153. Association for Computational Linguistics. Available at: https://doi.org/10.18653/v1/2023.emnlp-main.506.

Berkeley FrameNet. Accessed at: https://framenet.icsi.berkeley.edu/. (22 September

2025)

Bowman, S.R., Angeli, G., Potts, C. & Manning, C.D. (2015). A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, pp. 632–642. Association for Computational Linguistics.

Camacho-Collados, J. & Navigli, R. (2016). Find the word that does not belong: A Framework for an Intrinsic Evaluation of Word Vector Representations. In *Proceedings of the 1st Workshop on Evaluating Vector Space Representations for NLP*, pp. 43–50. Association for Computational Linguistics.

Comșa, I., Eisenschlos, J. & Narayanan, S. (2022). MiQA: A Benchmark for Inference on Metaphorical Questions. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 373–381. Available at: https://doi.org/10.18653/v1/2022.aacl-short.46.

Cruse, D.A. (1986). *Lexical Semantics*. Cambridge University Press.

Det Danske Sprog- og Litteraturselskab (2025). *Den Danske Ordbog.* Accessed at: https://ordnet.dk/ddo. (17 July 2025)

Einarsson, H., Simonsen, A. & Nielsen, D. S. (2025). In *Proceedings of the 1st Workshop on Nordic-Baltic Responsible Evaluation and Alignment of Language Models (NB-REAL 2025)*. Association for Computational Linguistics. Available at: https://aclanthology.org/2025.nbreal-1.0.

Hershcovich, D., Frank, S., Lent, H., de Lhoneux, M., Abdou, M., Brandl, S., Bugliarello, E., Cabello Piqueras, L., Chalkidis, I., Cui, R., Fierro, C., Margatina, K., Rust, P. & Søgaard, A. (2022). Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* pp. 6997–7013. Association for Computational Linguistics. Available at: https://doi.org/10.18653/v1/2022.acl-long.482.

Kilgarriff, A. (1997). I don't believe in word senses. *Computers and the Humanities*, 31(2), 91–113. Available at: https://doi.org/10.1023/A:1000583911091.

Nielsen, F. Å. & Hansen, L.K. (2017). Open semantic analysis: The case of word level semantics in Danish. In *Human Language Technologies as a Challenge for Computer Science and Linguistics*, pp. 415–419.

Nimb, S., Flörke, I., Olsen, S., Pedersen, B.S. & Sørensen, N.C.H. (2024a). COR.SEM, a New Formal Semantic Lexicon for Danish. In *Lexicography and Semantics: Proceedings of the XXI EURALEX International Congress,* Cavtat, Crotia.

Nimb, S., Lorentzen, H., T. Troelsgård, L. Theilgaard (2014). *Den Danske Begrebsordbog.* Det Danske Sprog-og Litteraturselskab.

Nimb, S., Sørensen, N.C.H. & Jensen, J. (2024b). Making Danish Thesaurus Data Available to Researchers – The WebDDB project. In *Lexicography and Semantics: Proceedings of the XXI EURALEX International Congress,* Cavtat, Crotia.

Nimb, S., Olsen, S., Pedersen, B. S., & Troelsgaard, T. (2022). A Thesaurus-based

Sentiment Lexicon for Danish: The Danish Sentiment Lexicon. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk & S. Piperidis (eds.) *Proceedings of the Language Resources and Evaluation Conference: LREC2022, Volume 2022*, pp. 2826–2832. Marseille. European Language Resources Association.

Nimb, S., Sørensen, N.H. & Troelsgård, T. (2018). From standalone thesaurus to integrated related words in the Danish Dictionary. In J. Čibej, V. Gorjanc, I. Kosem & S. Krek (eds.) *Proceedings from Euralex 2018*, pp. 915–923. Ljubljana, Slovenia. Znanstvena založba Filozofske fakultete Univerze v Ljubljani / Ljubljana University Press, Faculty of Arts.

Nimb, S. et al. (2017). From thesaurus to framenet. Electronic lexicography in the 21st century. *Proceedings of eLex 2017*, pp. 1–22. Brno: Lexical Computing CZ s.r.o. Available at: https://elex.link/elex2017/wp-content/uploads/2017/09/paper01.pdf.

Pedersen, B., Sørensen, N., Nimb, S., Hansen, D., Olsen, S. & Al-Laith, A. (2025). Evaluating LLM-Generated Explanations of Metaphors – A Culture-Sensitive Study of Danish. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pp. 470–479. University of Tartu Library. Available at: https://hdl.handle.net/10062/107190.

Pedersen, B. S., Sørensen, N.C.H., 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 *Proceedings of the 13th Language Resources and Evaluation Conference,* pp. 51–60. Marseille, France. European Language Resources Association.

Pedersen, B., Sørensen, N., Olsen, S., Nimb, S. & Gray, S. (2024). Towards a Danish semantic reasoning benchmark – Compiled from lexical-semantic resources for assessing selected language understanding capabilities of large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 16353–16363. ELRA and ICCL. Available at: https://aclanthology.org/ 2024.lrec-main.1421.

Pilehvar, M.T. & Camacho-Collados, J. (2019). WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Volume 1 (Long and Short Papers)*, pp. 1267–1273. Minneapolis, Minnesota: Association for Computational Linguistics.

Ruppenhofer, J., Ellsworth, M., Petruck, M.R.L., Johnson, C.R., Baker, C.F. & Scheffczyk, J. (2016). *FrameNet II: Extended theory and practice*. International Computer Science Institute. Available at: https://framenet.icsi.berkeley.edu/the_book.

Samuel, D. et al. (2023). NorBench – a benchmark for Norwegian language models. In

Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa), pp. 618–633. Tórshavn: University of Tartu Library. Available at: https://aclanthology.org/2023.nodalida-1.61.

Zhang, X., Li, S., Hauer, B., Shi, N. & Kondrak, G. (2023). Don't Trust ChatGPT when your Question is not in English: A Study of Multilingual Abilities and Types of LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7915–7927. Singapore. Association for Computational Linguistics. Available at: https://aclanthology.org/2023.emnlp-main.491/.

Better something than nothing: Analysis of GPT-4 performance in identifying Croatian proverbs

Nikola Bakarić

University of Applied Sciences Velika Gorica, Zagrebačka 5, Velika Gorica, Croatia
E-mail: nbakaric@vvg.hr

Abstract

The task of automatic detection of idiomatic expressions such as proverbs is an established problem in natural language processing. Before the advent of large language models, attempts were made to describe proverbs by modelling their syntactic structure (Rassi et al., 2014). Later, others employed contextual embeddings and neural networks to identify idioms (Škvorc et al., 2022) which is a task closely related to proverb detection.

This research effort aims to analyse the performance of the ChatGPT large language model (ChatGPT 4o) in the task of detecting proverbs and proverb-related expressions. As proverbs are often used in political discourse to underscore messages or augment arguments and points of view (G¡ndara, 2004), the research presented here will use the minutes of the Croatian parliament sessions made available by the Croatian parliamentary corpus ParlaMeter-hr (Dobranić et al., 2019) to build a list of proverbs occurring in contemporary discourse.

A list of 151 Croatian proverbs used in contemporary speech and texts was obtained from (Varga & Matovac, 2016) and other sources. Proverbs are mostly used as idiomatic expressions, with little variation. This fact was used to create a custom simple fuzzy search algorithm, which was then applied to a small section of the ParlaMeter-hr corpus to extract sentences which contain proverbs. The extracted list was further manually checked and verified. This simple search technique yielded 126 confirmed occurrences of sentences which contained proverbs.

The next step included prompting GPT-4o with a combination of prompts to determine its ability to detect proverbs, using both the chat and API interface. The prompts ranged from a very simple zero-shot to elaborate instructions with accompanying list of proverbs.

It was discovered that GPT-4o created a list of Croatian proverbs as a response to the chat based zero-prompt which contained only 12 items. Uploading the list of proverbs resulted in only 54% accuracy. API prompt returned better results, the zero-shot prompt reached 79% accuracy in under 5 minutes, while the most elaborate many-shot prompt using the curated list of proverbs reached 94% accuracy, but took over 120

minutes at an increased financial cost.

Keywords: large language model; GPT-4o; proverbs; automatic detection

References

Dobranić, F., Ljubešić, N., & Erjavec, T. (2019). Croatian parliamentary corpus ParlaMeter-hr 1.0 (Corpus, Text No. http://hdl.handle.net/11356/1209; Version 1.0). Slovenian language resource repository CLARIN.SI. Accessed at: https://www.clarin.si/repository/xmlui/handle/11356/1209.

G¡ndara, L. (2004). 'They That Sow the Wind...': Proverbs and Sayings in Argumentation. *Discourse & Society*, 15(2–3), pp. 345–359. Available at: https://doi.org/10.1177/0957926504041023.

Rassi, A. P., Baptista, J., & Vale, O. (2014). Automatic Detection of Proverbs and their Variants [Application/pdf]. *OASIcs, Volume 38, SLATE 2014*, 38, pp. 235–249. Available at: https://doi.org/10.4230/OASICS.SLATE.2014.235.

Škvorc, T., Gantar, P., & Robnik-Šikonja, M. (2022). MICE: Mining Idioms with Contextual Embeddings. *Knowledge-Based Systems*, 235, 107606. Available at: https://doi.org/10.1016/j.knosys.2021.107606.

Varga, M. A., & Matovac, D. (2016). KROATISCHE SPRICHWÖRTER IM TEST. *Proverbium: Yearbook of International Proverb Scholarship*, 33(1). Available at: https://hrcak.srce.hr/278302.

Woordpeiler: A New Tool for Visualizing and Analyzing Lexical Trends in Contemporary Dutch

Kris Heylen, Vincent Prins, Katrien Depuydt, Jesse de Does,

Laura van Eerten, Thomas Haga

Dutch Language Institute, Rapenburg 61, 2311 GJ Leiden (The Netherlands)
E-mail: kris.heylen@ivdnt.org, Vincent.Prins@ivdnt.org, Katrien.Depuydt@ivdnt.org,
Jesse.deDoes@ivdnt.org, Laura.vanEerten@ivdnt.org, Thomas.Haga@ivdnt.org

Abstract

Representative monitor corpora with detailed metadata offer a solid empirical basis for documenting lexical innovation and change (Kosem et al. 2021). However, continuously updated time-stamped textual data presents challenges for data management, lexicographic analysis, and visualization. Building on its existing corpus infrastructure, the Dutch Language Institute (INT) has developed *Woordpeiler* ("Word Pollster", https://woordpeiler.ivdnt.org/), an online application to (a) visualize and analyze word frequencies over time and (b) support the analysis of neologisms and lexical trends in Dutch since 2000.

As part of its mission to maintain a sustainable Dutch language infrastructure, INT developed the Corpus Hedendaags Nederlands (CHN), currently (September 2025) containing 4.3 billion tokens across 10.6 million documents. The corpus supports INT's lexicographic workflow and is available through CLARIN. Daily and yearly data from major Dutch-language newspaper publishers (in the Netherlands, Belgium, Suriname, and the Dutch Caribbean) is processed via an automated workflow. All data is converted into a unified TEI format, enriched with metadata (e.g. language variety) and linguistic annotation. Using INT's BlackLab system (de Does et al. 2017), the data is indexed and published as weekly (internal) or monthly (external) CHN updates.

While CHN users could already obtain word frequencies through BlackLab's query interface, *Woordpeiler* adds visualization and trend analysis tools. Frequency data for POS-tagged word forms, lemmas, and bigrams are exported to a PostgreSQL database optimized with TimeScaleDB. Through *Woordpeiler's* interface (Fig. 1), users can generate interactive graphs for words and bigrams to visualize and compare changes in absolute and relative frequencies across customizable time intervals (day, week, month, year). Wild cards can be used for searches and graphs can be filtered or split by language variety (Belgium, Netherlands, Suriname, Caribbean), with tooltips providing statistics and links to the underlying corpus data. In advanced search, users can refine searches by lemma, part of speech and newspaper (only internally). Graphs can be downloaded PNGs or shared through unique URLs.

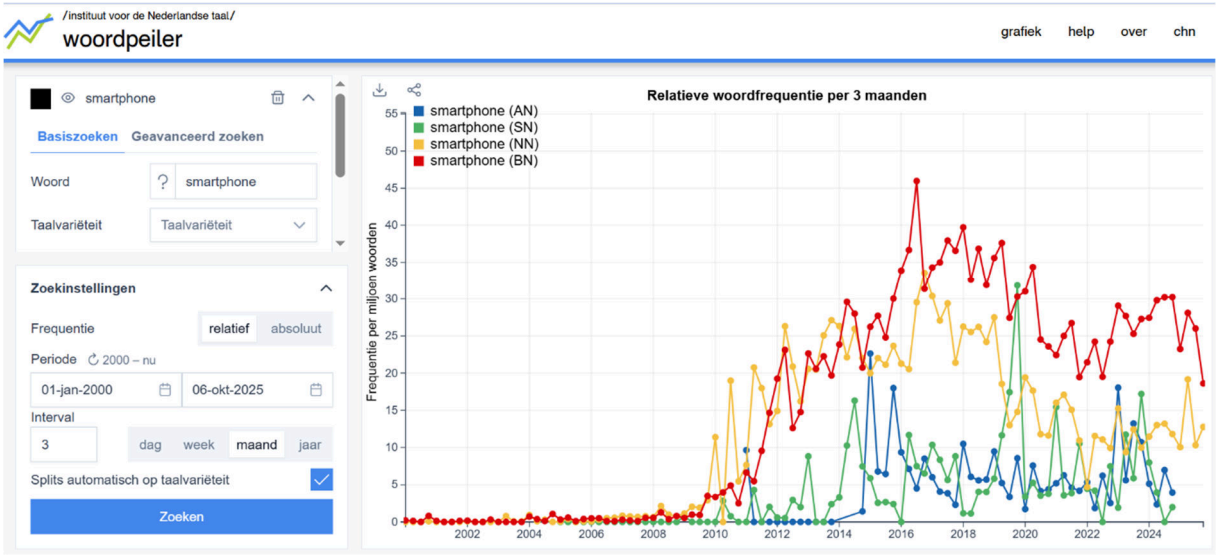


Figure 1: Basic search interface of *Woordpeiler,* showing the use of *smartphone* split by variety

A separate pane (Figure 2) offers additional trend analyses (currently only available internally). One function detects "trending" words or bigrams in a given interval using simple maths keyness (Kilgarriff 2009) relative to the preceding period. Users can adjust smoothing and also detect disappearing words via inverse keyness. A second function identifies new words or bigrams in a selected interval, optionally allowing a limited number of earlier nonce occurrences. Results appear as sortable, POS-filterable lists with accompanying frequency graphs.

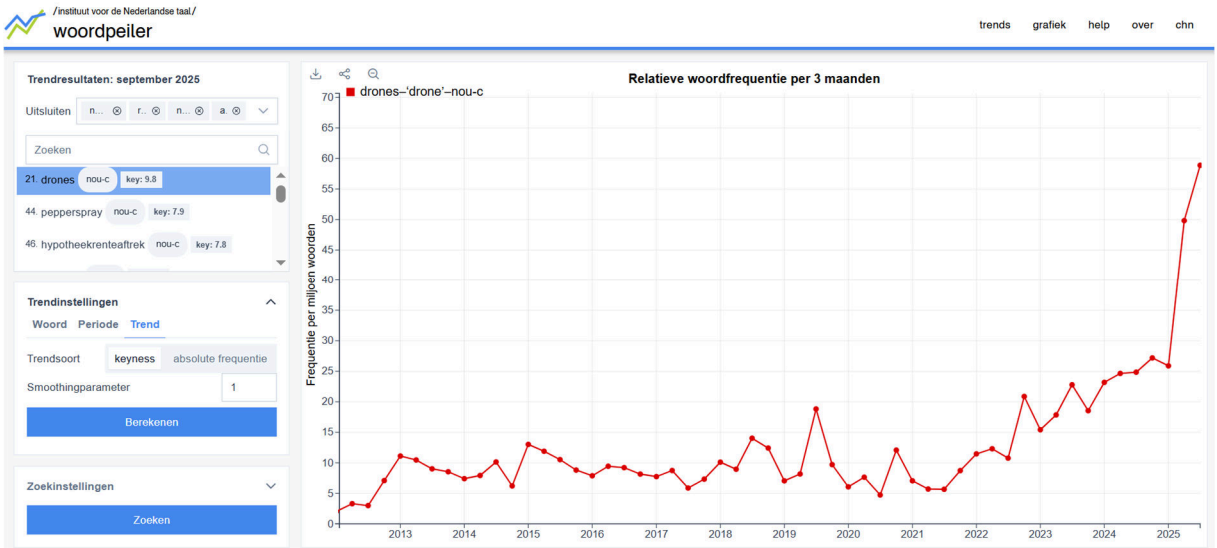


Figure 2: Trends interface of *Woordpeiler* showing *drones* as most trending noun in Sept. 2025

Woordpeiler and its database are fully integrated into INT's corpus-processing workflow, minimizing publication lags and ensuring quality control. The tool will support corpus-lexicographic work by adding validated frequency information to the central lexicon GiGaNT and improving workflows for identifying neologisms and out-of-dictionary words. Additionally, *Woordpeiler* serves science communication and outreach goals: it underpins a monthly and annual Woordpeiling ("Word Poll") shared via INT's website

and social media, and it is used in educational materials about language variation and change for secondary school students.

Keywords: lexical trends; neologism detection; corpus-based lexicography; monitor corpus; data visualization

References

De Does, J., Niestadt J. & Depuydt K. (2017). Creating research environments with BlackLab. In J. Odijk & A.van Hessen (eds.) *CLARIN in the Low Countries*, pp. 151–165.

Kilgarriff, A. (2009) Simple maths for keywords. In M. Mahlberg, V. González-Díaz, & C. Smith (eds.) *Proceedings of the Corpus Linguistics Conference CL2009.*

Kosem, I., Krek, S., Gantar, P., Arhar Holdt, Š. & Čibej, J. (2021). Language Monitor: Tracking the Use of Words in Contemporary Slovene. 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. 514–528.

AI- and Corpus-Based Strategies for Identifying Phraseme Constructions: A Pilot Study on Croatian Repetitive Constructions

Slobodan Beliga[1,2], Ivana Filipović Petrović[3]

[1] Faculty of Informatics and Digital Technologies, University of Rijeka, Radmile Matejčić 2, 51000 Rijeka, Croatia

[2] Center for Artificial Intelligence and Cybersecurity, University of Rijeka, Trg braće Mažuranića 10, 51000 Rijeka, Croatia

[3] Croatian Academy of Sciences and Arts, Trg Nikole Šubića Zrinskog 11, 10000 Zagreb, Croatia

E-mail: sbeliga@inf.uniri.hr, ifilipovic@hazu.hr

Abstract

The paper introduces a hybrid methodology for cross-linguistic identification of phraseme constructions, developed within the scope of a pilot study on Croatian repetitive constructions. The study explores how artificial intelligence and corpus technologies can be systematically combined to uncover functionally equivalent patterns across languages. The proposed strategy rests on three interdependent layers: (1) the AI layer, which harnesses large language models to generate candidate constructions, paraphrases, and corpus query formulations; (2) the corpus layer, which provides empirical validation through frequency data, authentic usage, and syntactic patterns; (3) and the human expert layer, which supervises prompt engineering, interprets outputs, and ensures linguistic adequacy. These layers operate in an iterative workflow, enabling dynamic interaction between computational and expert insights. The methodology is exemplified through the analysis of the German construction X über X 'X after X', for which the Croatian equivalent X za X-om (e.g., *dan za danom* 'day after day') is identified as structurally and semantically appropriate. The study compares outputs of two LLMs (GPT-4o and o3), revealing performance differences in idiomatic sensitivity. It also demonstrates how LLMs can assist in filtering corpus concordances to identify phraseologically valid examples. The study highlights both the strengths (e.g., scalability, reduced expert workload) and limitations (e.g., LLMs' sensitivity to prompt design and formal syntax) of the approach. It concludes that this layered strategy offers a viable path toward the semi-automatic processing of additional constructions and the development of multilingual phraseological resources.

Keywords: Repetitive phraseme constructions; Corpus Query Language; Large

Language Model; AI-assisted lexicography; Human-in-the-loop NLP

References

AbuMandour, W. (2024). Empowering bilingual lexicography with AI: The role of AImodel Triangulation Approach (AMTA) in dictionary design. In A. Inoue, N. Kawamoto & M. Sumiyoshi (eds.) *The 17th International Asian Association for Lexicography Conference*. Tokyo, Japan, pp. 70–83.

Beliga, S. & Filipović Petrović, I. (2024). Large Language Models Supporting Lexicography: Conceptual Organization of Croatian Idioms. In Š. Arhar Holdt & T. Erjavec (eds.) *Proceedings of the Conference on Language Technologies and Digital Humanities*. Ljubljana: Institute of Contemporary History, pp. 23–46.

Beliga, S. & Filipović Petrović, I. (2025). Can AI understand Croatian idioms? Assessing large language models in lexicographic tasks. *Prispevki za novejšo zgodovino*. Accepted for publication.

Chen, L., Dao, H.L. & Do-Hurinville, D.T. (2024). AI empowerment: Where are we in the automation of lexicography? A metaphraseographic study. In A. Inoue, N. Kawamoto & M. Sumiyoshi (eds.) *The 17th International Asian Association for Lexicography Conference*. Tokyo, Japan, pp. 90–98.

Davies, M. (2025). Integrating AI / LLMs into English-Corpora.org. Available at: https://www.english-corpora.org/ai-llms/. (15 June 2025)

Dobrovol'skij, D. (2018). Sind Idiome Konstruktionen? In K. Steyer (ed.) *Sprachliche Verfestigung. Wortverbindungen, Muster, Phrasem-Konstruktionen*, number 79 in Studien zur deutschen Sprache. Leibniz-Institut für Deutsche Sprache (IDS) [Zweitveröffentlichung], pp. 11–23.

Dobrovol'skij, D. (2011). Phraseologie und Konstruktionsgrammatik. In A. Lasch & A. Ziem (eds.) *Konstruktionsgrammatik III: Aktuelle Fragen und Lösungsansätze*. Stauffenburg, pp. 111–130.

Dobrovol'skij, D. (2022). *Deutsche Phrasem-Konstruktion [X hin, X her] in kontrastiver Sicht: eine korpusbasierte Analyse*. Berlin, Boston: De Gruyter, pp. 225–246. URL Available at: https://doi.org/10.1515/9783110770209-009.

Filipović Petrović, I. & Beliga, S. (2024). Lexicographic Treatment of Idioms and Large Language Models: What Will Rise to the Surface? In S. Krek (ed.) *Book of Abstracts of the Workshop Large Language Models and Lexicography*. Ljubljana: Centre for language resources and technologies, University of Ljubljana, pp. 12–16.

Filipović Petrović, I., López Otal, M. & Beliga, S. (2024). Croatian Idioms Integration: Enhancing the LIdioms Multilingual Linked Idioms Dataset. In N. Calzolari, M.Y. Kan, V. Hoste, A. Lenci, S. Sakti & N. Xue (eds.) *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evalu ation (LREC-COLING 2024)*. Torino, Italia: ELRA and ICCL, pp. 4106–4112. Available at: https://aclanthology.org/2024.lrec-main.366/.

Gantar, A. (2024). Formulisanje riječničkih definicija pomoću umjetne inteligencije na primjeru slovenskih frazeoloških jedinica. In S. Marjanović (ed.) *Moderni riječnici u funkciji prosečnog korisnika: stari problemi, suvremeni pravci i novi izazovi*, volume 1. Beograd: Filološki fakultet Univerziteta u Beogradu, pp. 151–157.

Goldberg, A.E. (1995). *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago and London: The University of Chicago Press.

Goldberg, A.E. (2019). *Explain Me This: Creativity, Competition and the Partial Productivity of Constructions*. Princeton and Oxford: Princeton University Press.

Hohenhaus, P. (2004). Identical Constituent Compounding – a Corpus-based Study. *Folia Linguistica*, 38(3-4), pp. 297–332. Available at: https://doi.org/10.1515/flin.2004.38.3-4.297.

Horn, L.R. (2018). *The lexical clone: Pragmatics, prototypes, productivity*. Berlin, Boston: De Gruyter Mouton, pp. 233–264. Available at: https://doi.org/10.1515/9783110592498-010.

Li, S., Chen, J., Yuan, S.,Wu, X., Yang, H., Tao, S. & Xiao, Y. (2024). Translate Meanings, Not Just Words: IdiomKB's Role in Optimizing Idiomatic Translation with Language Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17), pp. 18554–18563. Available at: https://ojs.aaai.org/index.php/ AAAI/article/view/29817.

Ljubešić, N. & Kuzman, T. (2024). CLASSLA-web: Comparable Web Corpora of South Slavic Languages Enriched with Linguistic and Genre Annotation. In N. Calzolari, M.Y. Kan, V. Hoste, A. Lenci, S. Sakti & N. Xue (eds.) *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Torino, Italia: ELRA and ICCL, pp. 3271–3282. Available at: https://aclanthology.org/2024.lrec-main.291/.

Ljubešić, N., Rupnik, P. & Kuzman, T. (2024). Croatian web corpus CLASSLA-web.hr 1.0. Available at: http://hdl.handle.net/11356/1929. Slovenian language resource repository CLARIN.SI.

Mellado Blanco, C. (2022). Phraseology, patterns and Construction Grammar: An introduction. In C. Mellado Blanco (ed.) *Productive Patterns in Phraseology and Construction Grammar: A Multilingual Approach*. Berlin, Boston: De Gruyter, pp. 1–26. Available at: https://doi.org/10.1515/9783110520569-001.

Mellado Blanco, C. (2023). From idioms to semi-schematic constructions and vice versa: The case of [a un paso de X]. In I. Hennecke & E. Wiesinger (eds.) *Constructions in Spanish*. John Benjamins Publishing Company, pp. 103–128. Available at: https://doi.org/10.1075/cal.34.05mel.

OpenAI (2024). GPT-4o System Card. Available at: https://arxiv.org/abs/2410.21276.

OpenAI (2025). o3-mini System Card. OpenAI technical documentation. Available at: https://cdn.openai.com/o3-mini-system-card-feb10.pdf.

Pavlova, A. (2024). Äquivalenz bei Übersetzung von Phrasemkonstruktionen. In A. Gondek, A. Jurasz, M. Kałasznik, P. Staniewski, J. Szczęk & A. Kamińska (eds.) *Interkulturelles und Interdisziplinäres in der Phraseologie und Parömiologie. Bd. II*. Hamburg: Verlag Dr. Kovač, pp. 159–178.

Pavlova, A., Naiditch, L. & Pöppel, L. (2022). Est' kreativnost' i kreativnost'. O granjah i granicah kreativnosti v oblasti frazeologizmov-konstrukcij. *Anzeiger für Slavische Philologie*, 50(1), pp. 181–204.

PhKW (2024). PhKWB Repository: Artikel. Available at: https://github.com/PhKW/PhKWB/tree/main/Artikel. Last updated: November 3, 2024; Accessed: April 1, 2025.

Piunno, V. (2022). Coordinated constructional intensifiers: patterns, function and productivity. In C. Mellado Blanco (ed.) *Productive Patterns in Phraseology and Construction Grammar: A Multilingual Approach*. Berlin, Boston: De Gruyter, pp. 133–164.

Shalevska, E., Kostadinovska-Stojchevska, B., Janusheva, V., Janusheva, M., Stojanoska, M. & Talevska, M. (2025). AI IN NONLITERAL LANGUAGE TRANSLATION: TRANSLATING MACEDONIAN PROVERBS AND IDIOMS. *International journal of Education Teacher*, 29, pp. 60–66. Available at: https://www.ijeteacher.com/index.php/ijet/article/view/83.

Sřrensen, N.H. & Nimb, S. (2025). The Danish Idiom Dataset: A collection of 1000 Danish idioms and fixed expressions. In H. Einarsson, A. Simonsen & D.S. Nielsen (eds.) *Proceedings of the 1st Workshop on Nordic-Baltic Responsible Evaluation and Alignment of Language Models (NB-REAL 2025)*. Tallinn, Estonia: The University of Tartu Library, pp. 55–63. Available at: https://aclanthology.org/2025.nbreal-1.5/.

Steyer, K. (2013). *Usuelle Wortverbindungen. Zentrale Muster des Sprachgebrauchs aus korpuslinguistischer Sicht*. Number 65 in Studien zur deutschen Sprache. Tübingen: Narr Francke Attempto Verlag.

Ziem, A. (2018). Construction Grammar meets Phraseology: eine Standortbestimmung. *Linguistik Online*, 90(3). Available at: https://doi.org/ 10.13092/lo.90.4316.

Exploring the power of generative artificial intelligence for automatic term extraction from small samples

Lena De Pourcq, Marie Grégoire, Leonardo Zilio

UCLouvain, Louvain-la-Neuve, Belgium

E-mail: {lena.depourcq,marie.gregoire}@student.uclouvain.be, leonardo.zilio@uclouvain.be

Abstract

This study explores the use of several chatbots based on recent generative large language models for automatic term extraction (ATE) from smaller text samples. The samples were selected from three domains: board games, ice hockey, and kitesurfing; and they cover three languages: English, French, and Portuguese. We used four prompting strategies: zero shot, one shot, few shots, and few shots with context. A single prompt with placeholders for language, domain and examples (when available) was used for all settings, and, in the case of French and Portuguese, we tested the ATE prompt in English and in the respective language. Results were calculated in terms of f-measure, and we further tested the best models with five consecutive runs to calculate a mean fmeasure and a standard deviation. No clear best system was verified for the task. Each of the domains and languages had different best systems. In terms of prompting strategy, more information did not always lead to better results, as zero-shot and one-shot attempts had the best results in several scenarios. The main contribution of the study is an overview of the ATE capacity of several chatbot systems across multiple scenarios.

Keywords: automatic term extraction; ATE; chatbots; generative artificial intelligence; GenAI

References

Ammon, U., Bickel, H. & Lenz, A.N. (eds.) (2016). *Variantenwörterbuch des Deutschen: Die Standardsprache in Österreich, der Schweiz, Deutschland, Liechtenstein, Luxemburg, Ostbelgien und Südtirol sowie Rumänien, Namibia und Mennonitensiedlungen*. Berlin/Boston: De Gruyter Mouton, 2 edition.

Andrade, J.C.B., Otálvaro, C.M.M., Jaramillo, C.M.Z. & Ríos, A.M. (2023). Approaches, tools, algorithms, and methods for automatic term extraction: A systematic literature mapping. *Research Square, preprint*.

Anthony, L. (2005). AntConc: design and development of a freeware corpus analysis toolkit for the technical writing classroom. In *IPCC 2005. Proceedings.*

International Professional Communication Conference, 2005. IEEE, pp. 729–737.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33, pp. 1877–1901.

Chun, Y., Kim, M., Kim, D., Park, C. & Lim, H. (2025). Enhancing Automatic Term Extraction with Large Language Models via Syntactic Retrieval. Available at: https://doi.org/10.48550/arXiv.2506.21222.

Costa, H., Zaretskaya, A., Corpas Pastor, G. & Seghiri Domínguez, M. (2016). Nine terminology extraction Tools: Are they useful for translators? *Multilingual*, 27(3), pp. 14–20.

De Paiva, V., Gao, Q., Kovalev, P. & Moss, L.S. (2023). Extracting Mathematical Concepts with Large Language Models. In *CEUR Workshop Proceedings*, pp. 1–13.

Di Nunzio, G.M., Marchesin, S. & Silvello, G. (2023). A systematic review of Automatic Term Extraction: What happened in 2022? *Digital Scholarship in the Humanities*, 38(Supplement_1), pp. i41–i47.

Giguere, J. (2023). Leveraging large language models to extract terminology. In *Proceedings of the First Workshop on NLP Tools and Resources for Translation and Interpreting Applications*, pp. 57–60.

Hellrich, J. & Hahn, U. (2016). An Assessment of Experimental Protocols for Tracing Changes in Word Semantics Relative to Accuracy and Reliability. In *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*. Berlin, Germany: Association for Computational Linguistics, pp. 111–117. Available at: http://anthology.aclweb.org/ W16-2114.

Kerremans, D., Stegmayr, S. & Schmid, H.J. (2011). The NeoCrawler: identifying and retrieving neologisms from the internet and monitoring ongoing change. In *Current Methods in Historical Semantics*. Berlin/Boston: De Gruyter, pp. 59–96. Available at: http://www.degruyter.com/view/books/9783110252903/9783110252903.59/9783110252903.59.xml.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. & Suchomel, V. (2014). The sketch engine. *Lexicography*, 1(1), pp. 7–36.

Kulkarni, M., Mahata, D., Arora, R. & Bhowmik, R. (2022). Learning Rich Representation of Keyphrases from Text. In M. Carpuat, M.C. de Marneffe & I.V. Meza Ruiz (eds.) *Findings of the Association for Computational Linguistics: NAACL 2022*. Seattle, United States: Association for Computational Linguistics, pp. 891–906.

Martínez-Cruz, R., López-López, A.J. & Portela, J. (2025). Chatgpt vs state-of-the-art models: a benchmarking study in keyphrase generation task. *Applied Intelligence*, 55(1), pp. 50.

Oliver, A. (2017). A system for terminology extraction and translation equivalent detection in real time: Efficient use of statistical machine translation phrase tables. *Machine Translation*, 31(3), pp. 147–161.

Pazienza, M.T., Pennacchiotti, M. & Zanzotto, F.M. (2005). Terminology extraction:

an analysis of linguistic and statistical approaches. In *Knowledge mining: Proceedings of the NEMIS 2004 final conference*. Springer, pp. 255–279.

Ramisch, C., Villavicencio, A. & Boitet, C. (2010). mwetoolkit: A framework for multiword expression identification. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Valletta, Malta: European Language Resources Association (ELRA), pp. 662–669.

Šajatović, A., Buljan, M., Šnajder, J. & Bašić, B.D. (2019). Evaluating automatic term extraction methods on individual documents. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pp. 149–154.

Salinas, A. & Morstatter, F. (2024). The Butterfly Effect of Altering Prompts: How Small Changes and Jailbreaks Affect Large Language Model Performance. In *Findings of the Association for Computational Linguistics ACL 2024*, pp. 4629–4651.

Schmidlin, R. (2011). *Die Vielfalt des Deutschen: Standard und Variation - Gebrauch, Einschätzung und Kodifizierung einer plurizentrischen Sprache*. Berlin: De Gruyter.

Siu, S.C. (2023). ChatGPT and GPT-4 for professional translators: Exploring the potential of large language models in translation. Available at: https://ssrn.com/abstract=4448091.

Song, M., Jiang, H., Shi, S., Yao, S., Lu, S., Feng, Y., Liu, H. & Jing, L. (2023). Is chatgpt a good keyphrase generator? a preliminary study. Available at: https://doi.org/10.48550/arXiv.2303.13001.

Südtiroler Kulturinstitut (2017). Sprachstelle im Südtiroler KULTURinstitut. Available at: http://www.kulturinstitut.org/hauptnavigation/sprachstelle.html.

Tran, H.T.H., Martinc, M., Caporusso, J., Doucet, A. & Pollak, S. (2023). The recent advances in automatic term extraction: A survey. Available at: https://doi.org/10.48550/arXiv.2301.06767.

Vidal Sabanés, L. & da Cunha, I. (2025). AI as a resource for the clarification of medical terminology: An analysis of its advantages and limitations. *Terminology*, 31(1), pp. 37–71.

Wang, S., Sun, X., Li, X., Ouyang, R., Wu, F., Zhang, T., Li, J. & Wang, G. (2023). GPT-NER: Named Entity Recognition via Large Language Models. Available at: https://doi.org/10.48550/arXiv.2304.10428.

Webber, W., Moffat, A. & Zobel, J. (2010). A similarity measure for indefinite rankings. *ACM Transactions on Information Systems*, 28(4), pp. 1–38.

Xu, K., Feng, Y., Li, Q., Dong, Z. & Wei, J. (2025). Survey on terminology extraction from texts. *Journal of Big Data*, 12(1), pp. 1–40.

Zilio, L., Paraguassu, L.B., Hercules, L.A.L., Ponomarekano, G.L., Berwanger, L.P. & Finatto, M.J.B. (2020). A lexical simplification tool for promoting health literacy. In *1st Workshop on Tools and Resources to Empower People with REAding DIfficulties*.

Lexicom at 25: reflections on the changing world of lexicography and language technology

Michael Rundell[1], Miloš Jakubíček[1,2], Vojtěch Kovář[1,2],

Ondřej Matuška[1], Michal Cukr[1]

[1] Lexical Computing, Czechia & United Kingdom
[2] Natural Language Processing Centre, Masaryk University, Brno, Czechia
E-mail: michael.rundell@gmail.com, firstname.lastname@sketchengine.eu

Abstract

In this paper we show how the academic content and computational tools featured in Lexicom form a parallel history of the last 25 years of innovation in lexicography. Lexicom is a 5-day intensive workshop offering handson training in corpus-based dictionary creation, from collecting and annotating language data to publishing the final product. Since it was launched in 2001, by Sue Atkins, Adam Kilgarriff, and Michael Rundell, Lexicom has adapted (sometimes incrementally, sometimes substantially), to reflect ongoing developments in linguistic theory, corpus tools, and NLP. Lexicom's curriculum integrates theoretical grounding with practical tasks such as corpus analysis, regular expressions, word sense disambiguation, and definition-writing. It provides an introduction to all of the key components of dictionary-creation and to the current state of the art in our field. The lexicographic landscape has seen transformative changes during Lexicom's 25-year lifetime. In 2001, corpora were relatively small even for well-resourced languages and non-existent for others; querying tools were quite basic; and the end-product was almost invariably a printed book. We now use billion-word corpora and sophisticated software to produce mainly digital dictionaries. Lexicom has mirrored these shifts, most recently incorporating AI and large language models. Amid all these dramatic changes, some constants in the dictionary-making process remain, and Lexicom continues to serve as both a reflection of and a guide through this ongoing evolution.

Keywords: dictionary; lexicography; Lexicom workshop; NLP, Sketch Engine; Post-editing lexicography; Large Language Models; teaching lexicography

References

Atkins, B.T.S. & Rundell, M. (2008). *The Oxford Guide to Practical Lexicography.* Oxford: Oxford University Press.

De Schryver, G.-M. (2023). Generative AI and Lexicography: the Current State of the

Art using ChatGPT. *International Journal of Lexicography*, 36(4), pp. 355–387.

Jakubíček, M., Měchura, M., Kovář, V. & Rychlý, P. (2018). Practical Post-Editing Lexicography with Lexonomy and Sketch Engine. In *Proceedings of the XVIII EURALEX Congress*. Ljubljana, Slovenia: Ljubljana University Press, Faculty of Arts, pp. 65–67. Available at: http://anthology.aclweb.org/W16-2114.

Jakubíček, M. & Rundell, M. (2023). The end of lexicography? Can ChatGPT outperform current tools for post-editing lexicography? *Electronic lexicography in the 21st century. Proceedings of the eLex 2023 conference*, pp. 518–533.

Kilgarriff, A. & Tugwell, D. (2001). WASP-Bench: an MT Lexicographers' Workstation Supporting State-of-the-art Lexical Disambiguation. In *Proceedings of Machine Translation Summit VIII*. Santiago de Compostela, Spain, pp. 187–190. Available at: https://kilgarriff.co.uk/publications.htm.

Klosa-Kückelhaus, A. & Tiberius, C. (2025). The Lexicographic Process Revisited. *International Journal of Lexicography*, 38(1), pp. 1–12.

Rundell, M. (2001). Teaching lexicography, or training lexicographers. *Kernerman Dictionary News*, 9, pp. 6–7. Available at: https://lexicala.com/wp-content/uploads/kdn9_2001_Teaching_lexicography_or_training_lexicographers_MR.pdf.

Rundell, M. (2023). Automating the creation of dictionaries: are we nearly there? In *Asialex 2023 Proceedings*. Seoul, South Korea, pp. 9–17.

Rundell, M., Jakubíček, M. & Kovář, V. (2020). Technology and English Dictionaries. In S. Ogilvie (ed.) *The Cambridge Companion to English Dictionaries*. Cambridge University Press, pp. 18–30.

Rundell, M. & Kilgarriff, A. (2011). Automating the creation of dictionaries: where will it all end? In F. Meunier, S. De Cock, G. Gilquin & M. Paquot (eds.) *A Taste for Corpora. A Tribute to Professor Sylviane Granger*. Benjamins, pp. 257–281.

Rychlý, P. (2007). Manatee/Bonito – A Modular Corpus Manager. In *First Workshop on Recent Advances in Slavonic Natural Language Processing, RASLAN 2007*. Brno: Masaryk University, pp. 65–70. Available at: https://nlp.fi.muni.cz/raslan/2007/.

Scott, M. (1999). *Wordsmith Tools Version 3*. Oxford: Oxford University Press.

Silva, P. (1998). Report on the SALEX '97 Lexicographical Training Course, 15–27 September 1997. *Lexikos*, 8, pp. 282–288.

Bridging human and AI perspectives: semantic annotation of generic nouns in German

Iván Arias-Arias[1], Elena Martín-Cancela[2]

[1] Instituto da Lingua Galega, Universidade de Santiago de Compostela, Praza da Universidade 4, 15782 Santiago de Compostela, Galicia

[2] Universidade da Coruña, Rúa Lisboa 7, Campus da Zapateira, 15008 A Coruña, Galicia

E-mail: ivanarias.arias@usc.gal, elena.martin@udc.es

Abstract

Generic nouns such as *Sache* and *Ding* pose a challenge for semantic annotation due to their referential underspecification and context-dependent meaning. Although frequently classified under categories like {artefact} or {object}, their actual referents often belong to abstract or cognitive domains, as in *Der Placeboeffekt ist eines der faszinierendsten Dinge in der Welt der Medizin*. Drawing on valency grammar, this study shows that these nouns activate different argument structures depending on their syntagmatic environment, reflecting semantic flexibility and combinatorial variability. Lexical databases such as GalNet or GermaNet frequently assign multiple synsets to these nouns, illustrating their ontological ambiguity. This paper examines whether large language models (LLMs) can replicate this nuanced classification. Using a gold standard corpus annotated by linguists, we implement a two-step prompting strategy —supplying LLMs with predefined semantic tags and contextual windows— to test their performance. The results underscore the limitations of current LLMs in dealing with the lexical underspecification of generic nouns, even when provided with an extended context window. These findings contribute to ongoing discussions on the automation of semantic tagging and point to meaningful ways in which AI systems can complement human expertise in natural language processing tasks.

Keywords: automatic semantic annotation; generic nouns; large language models; lexicological information systems; valency grammar

References

Alonso Ramos, M. (2023). El papel de ChatGPT como lexicógrafo. In C. Garriga Escribano et al. (eds.) *Lligams: Textos dedicats a Maria Bargalló Escrivŕ*. Tarragona: Publicacions Universitat Rovira i Virgili, pp. 15–27.

Anthropic (2025). Claude 4.0 Sonnet. Accessed at: https://claude.ai/.

Arias-Arias, I. (2025). Nuevas vías para la desambiguación en frases nominales en

alemán: fundamentos metodológico-lingüísticos para el desarrollo de una herramienta de anotación semántica (semi)automática. *Círculo de lingüística aplicada a la comunicación*. Forthcoming.

Arias-Arias, I., Domínguez Vázquez, M.J. & Valcárcel Riveiro, C. (2024). Der Effizienzund Intelligenzbegriff in der Lexikographie und künstlichen Intelligenz: kann ChatGPT die lexikographische Textsorte nachbilden? *Lexikos*, 34(1), pp. 51–76.

Berlin-Brandenburgische Akademie der Wissenschaften (2025). DWDS – Digitales Wörterbuch der deutschen Sprache. Accessed at: https://www.dwds.de/.

Bhattacharjee, A., Moraffah, R., Garland, J. & Liu, H. (2024). Zero-shot LLM-guided Counterfactual Generation: A Case Study on NLP Model Evaluation. Available at: https://doi.org/10.48550/arXiv.2405.04793.

Domínguez Vázquez, M.J. (2011). *Kontrastive Grammatik und Lexikographie: spanischdeutsches Wörterbuch zur Valenz des Nomens*. Munich: Iudicum.

Domínguez Vázquez, M.J. (2022). Estructura argumental del nombre: generación automática. *Signos: estudios de lingüística*, 55(119), pp. 732–761.

Domínguez Vázquez, M.J., Valcárcel Riveiro, C. & Bardanca Outeiriño, D. (2021). Portlex lexical ontology. Ontología léxica. Accessed at: http://portlex.usc.gal/ontologia/.

Dudenredaktion (2025). Duden online. Accessed at: https://www.duden.de/.

Engel, U. (2004). *Deutsche Grammatik: Neubearbeitung*. Munich: Iudicum.

Enis, M. & Hopkins, M. (2024). From LLM to NMT: Advancing Low-Resource Machine Translation with Claude. Available at: https://doi.org/10.48550/arXiv.2404.13813.

Google (2025). Gemini 2.5 Pro. Accessed at: https://gemini.google.com/.

Gómez Guinovart, X. & Solla Portela, M. (2019). GalNet. WordNet 3.0 do galego. Accessed at: https://ilg.usc.gal/galnet/.

Gödeke, L., Barth, F., Dönicke, T.,Weimer, A.M., Varachkina, H., Gittel, B., Holler, A. & Sporleder, C. (2022). Generalisierungen als literarisches Phänomen. Charakterisierung, Annotation und automatische Erkennung. *Zeitschrift für digitale Geisteswissenschaften*, 7.

Halliday, M.A.K. & Hasan, R. (1976). *Cohesion in English: Grammar and Text*. London: Longman.

Hamp, B. & Feldweg, H. (1997). GermaNet: a Lexical-Semantic Net for German. In *Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pp. 9–15.

Hinrichs, M., Lawrence, R. & Hinrichs, E. (2020a). Exploring and Visualizing Wordnet Data with GermaNet Rover. In *Proceedings of the CLARIN Annual Conference 2020*. Virtual Edition, pp. 32–26.

Hinrichs, M., Lawrence, R. & Hinrichs, E. (2020b). GermaNet Rover. Accessed at: https:// weblicht.sfs.uni-tuebingen.de/rover/.

Hölzner, M. (2007). *Substantivvalenz. Korpusgestützte Untersuchungen zu Argumentrealisierungen deutscher Substantive*. Tübingen: de Gruyter.

Institut für Deutsche Sprache (2025). DeReKo - Deutsches Referenzkorpus. Accessed at: https://cosmas2.ids-mannheim.de/cosmas2-web/.

Islam, R. & Ahmed, F. (2024). Gemini—the most powerful LLM: Myth or Truth. In *5th Information Communication Technologies Conference (ICTC)*. Nanjing, China, pp. 303–308.

Kolhatkar, V. & Hirst, G. (2014). Resolving Shell Nouns. In A. Moschitti et al. (eds.) *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar, pp. 499–510.

Kolhatkar, V., Zinsmeister, H. & Hirst, G. (2013). Interpreting anaphoric shell nouns using antecedents of cataphoric shell nouns as training data. In D. Yarowsky et al. (eds.) *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, Washington, pp. 300–310.

Landis, J.R. & Koch, G.G. (1977). The Measurement of Observer Agreement for Categorial Data. *Biometrics*, 33(1), pp. 159–174.

Mahlberg, M. (2005). *English general nouns: A corpus-theoretical approach*, *Volume 20 of Studies in Corpus Linguistics*. Amsterdam: John Benjamins.

Martín Gascueña, R. (2023). Diseño de una ontología de semántica léxica para los proyectos MultiGenera y MultiComb. In M.J. Domínguez Vázquez & C. Valcárcel Riveiro (eds.) *Desarrollo de aplicaciones para la generación automática del lenguaje: los recursos del portal lexicográfico Portlex (RILEX: Revista sobre investigaciones léxicas)*. Jaén: Revistas Científicas de la Universidad de Jaén, pp. 77–106.

Mel'čuk, I. (2015). *Semantics: From meaning to text*, *Volume 3*. Amsterdam: John Benjamins.

Mollica, F. (2010). *Korrelate im Deutschen und im Italienischen*. Frankfurt a.M.: Peter Lang.

Nasution, A.H. & Onan, A.A. (2024). ChatGPT Label: Comparing the Quality of Human-Generated and LLM-Generated Annotations in Low-Resource Language NLP Tasks. *IEEE Access*.

OpenAI (2025). ChatGPT-4o. Accessed at: https://chatgpt.com/.

Petukhova, K. & Kochmar, E. (2025). Intent Matters: Enhancing AI Tutoring with Fine-Grained Pedagogical Intent Annotation. Available at: https://doi.org/10.48550/ arXiv.2506.07626.

Pustejovsky, J. (1995). *The Generative Lexicon*. Cambridge MA: MIT Press.

Pustejovsky, J. & Batiukova, O. (2019). *The Lexicon*. Cambridge: Cambridge University Press.

Sántáné-Túri, A. (2020). *Die Selbständigkeit der Substantivvalenz*. Ph.D. thesis, University Szeged: SZTE Doktori Repozitórium.

Schmid, H.J. (2000). *English abstract nouns as conceptual shells: From corpus to cognition*. Berlin: De Gruyter Mouton.

Siddiky, M.N.A., Rahman, M.E., Hossen, M.F.B., Rahman, M.R. & Jaman, M.S. (2025). Optimizing AI language models: A study of ChatGPT-4 vs. ChatGPT-4o. *Preprints.org*.

Solla Portela, M.A. & Gómez Guinovart, X. (2015). Galnet: o WordNet do galego. Aplicacións lexicolóxicas e terminolóxicas. *Revista galega de filoloxía*, 16, pp. 169–201.

Sommerfeldt, K.E. & Schreiber, H. (1983). *Wörterbuch zur Valenz und Distribution der Substantive*. Berlin: de Gruyter.

Stefanowitsch, A. (2020). *Corpus linguistics: A guide to the methodology*. Berlin: Language Science Press.

Tarp, S. & Nomdedeu-Rull, A. (2023). Who has the last word? Lessons from using ChatGPT to develop an AI-based Spanish writing assistant. *Círculo de Lingüística Aplicado a la Comunicación*, 97, pp. 309–321.

Tiedemann, J. (2025). OPUS. Open Parallel Corpora. Accessed at: https://opus.nlpl.eu/.

Valcárcel Riveiro, C. & Pino Serrano, L. (2023). Application d'une méthodologie d'analyse des prédicats nominaux: l'exemple du lexčme MORT1. *Çédille: revista de estudios franceses*, 24, pp. 557–589.

Vossen, P. (1998). *EuroWordNet: A multilingual database with lexical semantic networks*. Dordrecht: Springer.

Wöllstein, A. & Dudenredaktion (2022). *Duden—Die Grammatik*. Mannheim: Dudenverlag.

Yu, D., Li, L., Su, H. & Fuoli, M. (2024). Assessing the potential of LLM-assisted annotation for corpus-based pragmatics and discourse analysis: The case of apologies. *International Journal of Corpus Linguistics*, 29(4), pp. 534–561.

Zhong, C., F., C., Liu, Q., Jiang, J., Wan, Z., Chu, C., Murawaki, Y. & Kuroshahi, S. (2024). Beyond English-Centric LLMs: What Language Do Multilingual Language Models Think In? Available at: https://doi.org/10.48550/arXiv.2408.10811.

A Corpus-Based Dictionary for the Endangered Megrelian Language

Irina Lobzhanidze, Rusudan Gersamia

Ilia State University, Kakutsa Cholokashvili Ave 3/5, Tbilisi 0179, Georgia
E-mail: irina_lobzhanidze@iliauni.edu.ge, rgersamia@iliauni.edu.ge

Abstract

This paper presents a corpus-based approach to compiling a bilingual Megrelian-English online dictionary. The Megrelian language belongs to the UNESCO Atlas of the World's Languages in Danger group of "increasingly endangered" languages, and faces a number of critical challenges, among them a lack of standardised resources, intergenerational transmission, and minimal digital presence. Unlike widely spoken languages equipped with pretrained models and various linguistic tools, "increasingly endangered" languages like Megrelian lack even basic NLP tools such as annotated corpora, PoS taggers, and morphological analysers. Moreover, the complexity of their grammar and phonology require special approaches that cannot simply be adapted from high-resource languages. To address these gaps, we developed an annotated corpus of contemporary Megrelian, consisting of 97691 tokens and 60959 types. It is based on data collected through fieldwork in Samegrelo, Georgia, from the years 2022 to 2025. The whole process was subdivided in two main stages: fieldwork conceptualization and data collection, followed by laboratory analysis and data processing.

The bilingual Megrelian-English dictionaries were developed in parallel, using the same dataset processed in Fieldworks Language Explorer (FLEx, 2024). This approach enabled the integration of corpus annotations into the dictionary entries. Following the principles described in Atkins & Rundell (2008), Gibbon & Van Eynde (2000), we used lexeme-based and root-based configurations, resulting in the creation of two online dictionaries, available online. The first dictionary is oriented toward the translation of individual words, while the second focuses on the translation of individual morphemes. In the first case, each lexical entry is supported by morphosyntactic information, phonetic transcription (IPA), glosses, and semantic descriptions. In the second case, the entries represent individual morphemes, providing not only glosses, but also information about their occurrences and links to their use in the corpus. The finalised data is available online through https://xmf.iliauni.edu.ge/.

The paper is subdivided into several parts: 1. Introduction, outlining the significance of Megrelian as part of the Kartvelian language family and introduces the project dedicated to the documentation of the Megrelian language; 2. Background and Data Collection, providing overviews the existing Megrelian dictionaries and represents the data collection stages; 3. Annotation and Corpus Development, describing the data

annotation and processing stages and giving information on corpus size, linguistic coverage, etc.; 4. The Dictionaries - Design and Generation, presenting the configurations for both the lexeme-based and morpheme-based dictionaries, and also thoroughly describing the export and converstion stages, oulining the linkage between the corpus and the dictionary entries, and; 5. Conclusions, Challenges and Future Works, which summarises the corpus-based lexicographic approach to the Megrelian language, provides a short description of the ongoing challenges, and describes future plans concerning the use and potential improvement of the data.

Keywords: Megrelian Language Corpus (MLC); endangered lexicography; bilingual dictionaries

References

Atkins, S.B.T. & Rundell, M. (2008). *The Oxford Guide to Practical Lexicography.* Oxford: Oxford University Press.

FLEx: SIL Feildworks Language Explorer (FLEx), Version 9.1. Accessed at: https://software.sil.org/fieldworks/. (20 July 2025)

Gibbon, D. & Van Eynde, Fr. (2000). *Lexicon Development for Speech and Language Processing.* London: Kluwer Academic Publishers.

Passive Vocabulary of Czech Native Speakers: A Statistical Approach

Marek Blahuš[1], Miloš Jakubíček[1,2], Vojtěch Kovář[1,3], František Kovařík[1,3]

[1] Lexical Computing, Brno, Czech Republic
[2] Faculty of Informatics, Masaryk University, Brno, Czech Republic
[3] Faculty of Arts, Masaryk University, Brno, Czech Republic
E-mail: firstname.lastname@sketchengine.eu

Abstract

This paper explores the theory of measuring vocabulary size, including the various methods that can be used and the parameters that have to be set. We have examined the experiments carried out on English and Dutch. Goulden et al. (1990) claims the average native speaker knows about 17,000 English base words (non-derived words). Keuleers et al. (2015) and Brysbaert et al. (2016) claim the average native speaker with secondary education knows about 42,000 headwords (lemmas). We have conducted an experiment similar to that of Keuleers and Brysbaert on Czech, with the input of 100,000 letter sequences from the wordlists of large web corpora. We assume the vocabulary size of Czech native speakers (as well as the vocabulary size of native speakers of any language) could be bigger, exceeding 57,000 (Czech) headwords, should we provide the participants with more inputs (150,000 sequences, or even more) or should we count the specialized terminology of their fields of interest.

Keywords: passive vocabulary; native speaker; manual annotation; semi-automatic dictionary drafting; Dictionary Express

References

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 *Proceedings of the 6th Biennial Conference on Electronic Lexicography*. Brno, Czech Republic: Lexical Computing CZ s.r.o., pp. 805–818. Available at: https://elex.link/elex2019/wp-content/uploads/2019/10/eLex-2019_Proceedings.pdf.

Blahuš, M., Cukr, M., Herman, O., Jakubíček, M., Kovář, V., Kraus, J., Medveď, M.,

Ohlídalová, V. & Suchomel, 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*. Brno, Czech Republic: Lexical Computing CZ s.r.o., pp. 613–637. Available at: https://elex.link/elex2023/wp-content/uploads/114.pdf.

Brysbaert, M., Stevens, M., Mandera, P. & Keuleers, E. (2016). How many words do we know? Practical estimates of vocabulary size dependent on word definition, the degree of language input and the participant's age. *Front. Psychol.*, 7, pp. 1116.

Diack, H. (1975). *Wordpower. Your Vocabulary and Its Measurement.* Paladin. St Albans.

Goulden, R., Nation, P. & Read, J. (1990). How large can a receptive vocabulary be? *Applied Linguistics*, 11, pp. 341–363. Available at: https://api.semanticscholar.org/CorpusID:145483070.

Henriksen, B. (1999). Three dimensions of vocabulary development. *Stud. Second Lang. Acquis.*, 21(2), pp. 303–317.

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), pp. 1665–1692. Available at: https://doi.org/10.1080/17470218.2015.1022560.

Kovařík, F. (2023). Semi-automatic Dictionary Creation for Czech. *Recent Advances in Slavonic Natural Language Processing (RASLAN 2023)*, 17. Available at: https://nlp.fi.muni.cz/raslan/raslan23.pdf.

Kovařík, F., Blahuš, M., Cukr, M., Jakubíček, M. & Kovář, V. (2024a). Dictionary Express: First Phases Rapid dictionary-making method for European, Asian and other languages. In A. Inoue, N. Kawamoto & M. Sumiyoshi (eds.) *AsiaLex 2024 Proceedings: Asian Lexicography- Merging cutting-edge and established approaches.* Tokyo: Toyo University, pp. 84–89. Available at: https://www.asialex.org/pdf/Asialex-Proceedings-2024.pdf.

Kovařík, F., Kovář, V. & Blahuš, M. (2024b). 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*. Cavtat: Institut za hrvatski jezik, pp. 336–344. Available at: https://euralex.jezik.hr/wp-content/uploads/2021/09/Euralex-XXI-proceedings_1st.pdf.

Merriam-Webster. (n.d.). F-HOLE. In *Merriam-Webster.com dictionary*. Accessed at: https://www.merriam-webster.com/dictionary/f-hole. (28 June 2025)

Měchura, M. (2017). Introducing Lexonomy: an Open-source Dictionary Writings and Publishing System. In I. Kosem, C. Tiberius, M. Jakubíček, J. Kallas, S. Krek & V. Baisa (eds.) *Electronic lexicography in the 21st century: Proceedings of eLex 2017 conference*. Lexical Computing.

Pignot-Shahov, V. (2012). Measuring L2 Receptive and Productive Vocabulary Knowledge. *Language Studie Working Papers*, 4(II), pp. 37–45.

R.L.G. (2013). Lexical facts. Available at: https://www.economist.com/johnson/2013/05/29/lexical-facts.

Suchomel, V. (2018). csTenTen17, a Recent Czech Web Corpus. In A. Horák, P. Rychlý & A. Rambousek (eds.) *Proceedings of the Twelfth Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2018*. Brno: Tribun EU, pp. 111–123. Available at: https://nlp.fi.muni.cz/raslan/2018/paper10-Suchomel.pdf.

Těšitelová, M. (1987). *O češtině v číslech*. Academia, malá jazyková knižnice edition.

Webb, S. (2020). *The Routledge Handbook of Vocabulary Studies*. Routledge Handbooks. New York, NY.

Automating Adjectival Microstructures in Monolingual Dictionaries: A New Method Combining Embeddings and LLMs

Enikő Héja, László Simon, Veronika Lipp

ELTE Research Centre for Linguistics, Budapest, 1068 Benczúr u. 33. Hungary
E-mail: heja.eniko@nytud.hu, simon.laszlo@nytud.hu, lipp.veronika@nytud.hu

Abstract

Recent findings indicate that current large language models (LLMs) face difficulties in generating clear-cut, well-motivated definitions in a consistent way. This shortcoming is the consequence of their reliance on opaque data sources and their inherently unstable, non-deterministic outputs. In response, this research aims to develop an LLM-based methodology for producing adjectival microstructures in monolingual dictionaries in a way that is both more consistent and aligned with lexicographic standards. Building on the hypothesis that prompts enriched with contextual information can enhance definition quality, the study employs a graph-based, interpretable, and unsupervised method starting out from static adjectival embeddings. The approach has previously demonstrated the ability to formalize traditional lexical semantic relations, detect adjectival senses from corpus data, and identify the most salient nominal contexts for each sense. The ultimate goal is to integrate these results into practical lexicographic workflows and assess how LLMs, when properly guided, can support dictionary compilation.

Keywords: unsupervised sense detection; graphs; adjectival microstructure; LLMs

References

Ah-Pine, J., & Jacquet, G. (2009). Clique-based clustering for improving named entity recognition systems. In A. Lascarides et al. (eds.) *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pp. 51–59.

Apresjan, J.D. (1974). Regular polysemy. *Linguistics*, 12(142), pp. 5–32.

Haber, J. & Poesio, M. (2024). Polysemy—Evidence from linguistics, behavioral science, and contextualized language models. *Computational Linguistics*, 50(1), pp. 351–417. Available at: https://doi.org/10.1162/coli_a_00500.

Comprehensive Dictionary of Hungarian = Ittzés, N. (chief ed.) (2006–). *A magyar nyelv nagyszótára*, I–VIII. Budapest: MTA Nyelvtudományi Intézet.

Concise Dictionary of Hungarian = Pusztai, F. (ed.). (2003). *Magyar értelmező*

kéziszótár. Budapest: Akadémiai Kiadó.

Héja, E. & Ligeti-Nagy, N. (2022a). A proof-of-concept meaning discrimination experiment to compile a word-in-context dataset for adjectives—A graph-based distributional approach. *Acta Linguistica Academica*, 69(4), pp. 521–548.

Héja, E. & Ligeti-Nagy, N. (2022b). A clique-based graphical approach to detect interpretable adjectival senses in Hungarian. In D. Ustalov, Y. Gao, A. Panchenko, M. Valentino, M. Thayaparan, T.H. Nguyen, G. Penn, A. Ramesh & A. Jana (eds.) *Proceedings of TextGraphs-16: Graph-based methods for natural language processing,* pp. 35–43. ACL.

Héja, E., Ligeti-Nagy, N., Simon, L. & Lipp, V. (2023). An unsupervised approach to characterize the adjectival microstructure in a Hungarian monolingual explanatory dictionary. In 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. 150–167. Lexical Computing.

Héja, E., Gábor, K., Simon, L. & Lipp, V. (2024a). Graph-based detection of Hungarian adjectival meaning structures via monolingual static embeddings. In K.Š. Despot, A. Ostroški Anić, & I. Brač (eds.), *Lexicography and Semantics: Proceedings of the XXI EURALEX International Congress*, pp. 235–247. Zagreb: Institute of Croatian Language and Linguistics.

Héja, E., Gábor, K., Győrffy, A., Ligeti-Nagy, N., Simon, L. & Lipp, V. (2024b). Melléknevek disztribúciós és szemantikai mintázatai. In V. Lipp, N. Ligeti-Nagy & L. Simon (eds.) *Prószéky Gábor 70: PG70 – Ünnepi kötet*, pp. 44–51. Budapest: HUN-REN Nyelvtudományi Kutatóközpont.

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 (eLex 2023): Invisible lexicography. Proceedings of the eLex 2023 conference*, pp. 518–533. Lexical Computing.

Ježek, E. (2016). *The lexicon: An introduction.* Oxford University Press.

Kiefer, F. (2000). *Jelentéselmélet.* Budapest: Corvina.

Kiefer, F. (2008). A melléknevek szótári ábrázolásáról. In *Strukturális magyar nyelvtan 4: A szótár szerkezete.* Budapest: Akadémiai Kiadó, pp. 505–538.

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M. & Rychlý, P. (2008). GDEX: Automatically finding good dictionary examples in a corpus. In *Proceedings of the XIII EURALEX international congress*, Barcelona, pp. 425–432.

Miller, G.A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, 38(11), pp. 39–41.

Ostermann, Carolin. (2015). *Cognitive Lexicography: A New Approach to Lexicography Making Use of Cognitive Semantics*, Berlin, München, Boston: De Gruyter. Available at: https://doi.org/10.1515/9783110424164.

Rehurek, R. & Sojka, P. (2011). *Gensim—Python framework for vector space modeling.*

NLP Centre, Faculty of Informatics, Masaryk University, 3(2).

Š. Despot, K., Ostroški Anić, A. & Brač, I. (eds.) (2024). *Lexicography and Semantics: Proceedings of the XXI EURALEX International Congress 8–12 October 2024*. Zagreb: Institute of Croatian Language and Linguistics.

Svensén, Bo. (2009). *A Handbook of Lexicography: The Theory and Practice of Dictionary-Making*, Cambridge: Cambridge University Press.

Zinoviev, D. (2018). *Complex network analysis in Python: Recognize - Construct - Visualize - Analyze - Interpret*. The Pragmatic Bookshelf.

Automatic Non-recorded Sense Detection for Swedish through Word Sense Induction with fine-tuned Word-in-Context models

Dominik Schlechtweg[1], Emma Sköldberg[2], Shafqat Mumtaz Virk[2], James White[1], Simon Hengchen[3]

[1] University of Stuttgart
[2] University of Gothenburg
[3] Université de Genève & iguanodon.ai
E-mail: first.last@ims.uni-stuttgart.de, first.last@svenska.gu.se, first.last@unige.ch

Abstract

Finding non-recorded senses is important for dictionary maintenance, where using automatic methods helps reduce manual efforts. We use automatic Word Sense Induction (WSI) to compare recorded sense numbers among a sample of headwords in a comprehensive Swedish monolingual dictionary with induced sense numbers for the same words in a Swedish corpus. We propose this as a simple technique to find words to prioritize for post-hoc manual checks, which can be done in a simple Online-User-Interface bypassing the need for programming knowledge. We perform a thorough manual evaluation of the proposed methodology enabling us to show statistically that using automatic WSI increases the odds of finding non-recorded senses compared to a random selection of words. We further (i) evaluate predictions according to potential inclusion in the dictionary providing strong evidence for usefulness in practical lexicography, and (ii) analyze model predictions in-depth to point towards future improvements. We, finally, integrate lessons learned from our analysis into a large-scale prediction effort, providing the first high-quality large-scale WSI predictions for Swedish. These are a valuable resource for future research in Swedish lexicography.

Keywords: Non-recorded Sense Detection; Word Sense Induction; Word-in-Context; DURel; Swedish

References

Cassotti, P., Siciliani, L., de Gemmis, M., Semeraro, G. & Basile, P. (2023). Xl-lexeme: WiC pretrained model for cross-lingual lexical semantic change. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics.*

Association for Computational Linguistics, pp. 1577–1585.

Cheilytko, N. & von Waldenfels, R. (2024a). Semantic change and lexical variation in Ukrainian with vector representations and LLM. In S. Krek (ed.) *Book of Abstracts of the Workshop Large Language Models and Lexicography*, pp. 1–5.

Cheilytko, N. & von Waldenfels, R. (2024b). Word Embeddings for Detecting Lexical Semantic Change in Ukrainian. In K.Š. Despot, A. Ostroški Anić & I. Brač (eds.) *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress*, pp. 231–243.

Cook, P., Lau, J. H., Rundell, M., McCarthy, D. & Baldwin, T. (2013). A lexicographic appraisal of an automatic approach for detecting new word senses. In I. Kosem, J. Kallas, P. Gantar, S. Krek, M. Langemets, & M. Tuulik (eds.) *Proceedings of eLex 2013 conference*. Ljubljana/Tallinn: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut, pp. 49–65.

Cook, P., Lau, J.H., McCarthy, D. & Baldwin, T. (2014). Novel word-sense identification. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 1624–1635.

Erk, K. (2006). Unknown word sense detection as outlier detection. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*. New York City, USA: Association for Computational Linguistics, pp. 128–135.

Fedorova, M., Mickus, T., Partanen, N., Siewert, J., Spaziani, E. & Kutuzov, A. (2024). AXOLOTL'24 Shared Task on Multilingual Explainable Semantic Change Modeling. In N. Tahmasebi, S. Montariol, A. Kutuzov, D. Alfter, F. Periti, P. Cassotti & N. Huebscher (eds.) *Proceedings of the 5th Workshop on Computational Approaches to Historical Language Change*. Bangkok, Thailand: Association for Computational Linguistics, pp. 72–91.

Fisher, R.A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 222(594–604), pp. 309–368.

Grundy, V. & Rawlinson, D. (2015). The practicalities of dictionary production; planning and managing dictionary projects; training of lexicographers. In P. Durkin (ed.) *The Oxford Handbook of Lexicography*. Oxford: Oxford University Press, pp. 561–578.

Hammarstedt, M., Schumacher, A., Borin, L. & Forsberg, M. (2022). Sparv 5 User Manual.

Jana, A., Mukherjee, A. & Goyal, P. (2020). Network measures: A new paradigm towards reliable novel word sense detection. *Information Processing & Management*, 57(6), pp. 102173.

Kokosinskii, D., Kuklin, M. & Arefyev, N. (2024). Deep-change at AXOLOTL-24: Orchestrating WSD and WSI models for semantic change modeling. In *Proceedings of the 5th Workshop on Computational Approaches to Historical Language Change*. Bangkok, Thailand. Association for Computational Linguistics, pp. 168–179.

Kurtyigit, S., Park, M., Schlechtweg, D., Kuhn, J. & Schulte im Walde, S. (2021).

Lexical semantic change discovery. Available at: https://doi.org/10.48550/arXiv.2106.03111.

Lau, J. H., Cook, P., McCarthy, D., Newman, D. & Baldwin, T. (2012). Word sense induction for novel sense detection. In W. Daelemans (ed.) *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Avignon, France: Association for Computational Linguistics, pp. 591–601.

Lautenschlager, J., Sköldberg, E., Hengchen, S., Schlechtweg, D. (2024). Detection of Nonrecorded Word Senses in English and Swedish. Available at: https://arxiv.org/abs/2403.02285.

Nilsson, P. (2024). Report on the revision of the Swedish Academy Dictionary – and the search for "old neologisms". In K.Š. Despot, A. Ostroški Anić & I. Brač (eds.) *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress*. Zagreb: Institute of Croatian Language and Linguistics, pp. 507–522.

Nimb, S., Sřrensen, N.H. & Lorentzen, H. (2020). Updating the dictionary: Semantic change identification based on change in bigrams over time. *Slovenščina 2.0: empirične, aplikativne in interdisciplinarne raziskave*, 8(2), 112–138.

Pilehvar, M.T. & Camacho-Collados, J. (2019). WiC: the Word-in-Context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota. Association for Computational Linguistics, pp. 1267–1273.

Sander, P., Hengchen, S., Zhao, W., Ma, X., Sköldberg, E., Virk, S. & Schlechtweg, D. (2024). The DURel Annotation Tool: Using fine-tuned LLMs to discover non-recorded senses in multiple languages. In *Proceedings of the Workshop on Large Language Models and Lexicography at 21st EURALEX International Congress Lexicography and Semantics*.

Schlechtweg, D., McGillivray, B., Hengchen, S., Dubossarsky, H. & Tahmasebi, N. (2020). SemEval-2020 Task 1: Unsupervised lexical semantic change detection. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, Barcelona, Spain. Association for Computational Linguistics.

Schlechtweg, D., Virk, S., Sander, P., Sköldberg, E., Theuer Linke, L., Zhang, T., Tahmasebi, N., Kuhn, J. & Schulte Im Walde, S. (2024a). The DURel Annotation Tool: Human and Computational Measurement of Semantic Proximity, Sense Clusters and Semantic Change. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, pp. 137–149.

Schlechtweg, D., Zamora-Reina, F.D., Bravo-Marquez, F. & Arefyev, N. (2024b). Sense through time: diachronic word sense annotations for word sense induction and lexical semantic change detection. *Language Resources and Evaluation*, pp. 1–35.

Sköldberg, E., Virk, S., Sander, P., Hengchen, S. & Schlechtweg, D. (2024). Revealing semantic variation in Swedish using computational models of semantic proximity

– Results from lexicographical experiments. In K.Š. Despot, A. Ostroški Anić & I. Brač (eds.) *Proceedings of the 21st EURALEX International Congress Lexicography and Semantics*, pp. 169–182.

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. (Digital Linguistics, 3)*. De Gruyter, pp. 53–82.

SO: Svensk ordbok utgiven av Svenska Akademien ['The Contemporary Dictionary of the Swedish Academy']. (2021). 2nd edition. Retrieved May 15, 2024. Accessed at: https://svenska.se/. (15 May 2025)

SVT (Sweden Television) corpus. Retrieved May 15, 2024, from SprÍkbanken's word research platform Korp. Accessed at: https://spraakbanken.gu.se/korp.

Yadav, S. & Schlechtweg, D. (2025). XL-DURel: Finetuning Sentence Transformers for Ordinal Word-in-Context Classification. Available at: https://arxiv.org/abs/2507.14578.

Parsing of Explanatory dictionary

Iryna Ostapova[1], Yevhen Kupriianov[2], Mykyta Yablochkov[1]

[1] Ukrainian Lingua-Information Foundation of National Academy of Sciences of Ukraine,
3 Holosiivska avenue, 03039, Kyiv, Ukraine
[2] National Technical University "Kharkiv Polytechnic Institute",
2 Kyrpychova str., 61002, Kharkiv, Ukraine
E-mail: irinaostapova@gmai.com, eugeniokuprianov@gmail.com, gezartos@gmail.com

Abstract

The paper outlines technological and methodological ways to arrange the dictionary parsing process. The Spanish Dictionary (Diccionario de la lengua Española 23 ed. – DLE 23) website (https://dle.rae.es/) serves as a basis for the research. First of all, asthe most complex multi-parameter lexicographic frameworks, explanatory dictionaries of national languages are of the most interest because they offer the most comprehensive lexicographic description of a language, are produced by top experts (linguists and IT engineers), and offer numerous opportunities to fully utilize contemporary digital technologies.

Ultimately, our goal is to create a digital version of the Dictionary of Spanish that can be easily adjusted to the user's evolving demands using a built-in research toolbox. Toachieve it we started the project named as Virtual Lexicographic Laboratory of the Dictionary of Spanish (VLL DLE 23) is the title of the project.

The first step was to build up a formal model that would serve as a basis to elaborate parsing algorithm, XML schema, database schema and interfaces. The formal model of DLE 23 was built based on analyzing the structure of dictionary entries of the online version and the printed variant of DLE 23.

The second step is to create a lexicographic database. Since the dictionary entries have a strictly defined structure, it makes sense to represent them as classes in object-oriented programming languages with subsequent processing, editing and storage in explicit form. NoSQL databases (document-oriented databases) provide such apossibility. LiteDB database (http://www.litedb.org/) was chosen for our project.

The final stage of the trial version was creating a web application to work with the VLL DLE database The application was created on the basis of .Net Core 2.1 technology. A set of HTML, CSS templates and JavaScript Bootstrap scripts was used for convenience and modification of interface elements.

The DLE 23 VLL project is realized in two stages: 1) creation of a VLL pilot version to test specific technological solutions and clarify the structure of the dictionary entry;

2) development of a final application with a full-scale interface. Currently, the first stage has been completed. The pilot version demonstrates more possibilities for the user than the original online version of DLE 23. Streaming version of DLE 23 is available at https://svc2.ulif.org.ua/Dics/ResIntSpanish (captcha is used).

Further parameterization of dictionary entries was done in order to construct the pilot version of the VLL. A collection of parameters is associated with each headword: 1) headword variations; 2) headword structure; 3) headword type; 4) homonymy; 5) number of meanings; 6) number of word combinations, and some others. Each parameter was identified using the dictionary entry's HTML text as a baseline. To create a selection, the user can enter any combination of these parameters. Articles are shown in a manner akin to the original edition, and the HTML-formatted text is also displayed. Statistics are produced for every selection. Full-text search is an additional option that can be combined with parametric search. You can specify any line of HTML text as a search string.

Keywords: lexicographic system; lexicographic data model; data analysis;

lexicographic database; user interface

Using Large Language Models to Generate Distractors for Language Games

Iztok Kosem[1,2,3], Špela Arhar Holdt[1,2]

[1] Faculty of Arts, University of Ljubljana, Aškerčeva 2, 1000, Ljubljana, Slovenia
[2] Faculty of Computer and Information Science, University of Ljubljana, Večna pot 113, 1000, Ljubljana, Slovenia
[3] Institut "Jožef Stefan", Jamova cesta 39, 1000, Ljubljana, Slovenia
E-mail: Spela.ArharHoldt@ff.uni-lj.si, Iztok.Kosem@fri.uni-lj.si

Abstract

This paper presents two tasks involving large language models (LLMs)—Gemini-2.0-flash and GPT-4o—used to generate distractors (i.e., incorrect options) for synonym and collocation questions in a language game. The lexical data for both tasks was sourced from the *Digital Dictionary Database of Slovene* (DDDS). Prompts were initially tested on a sample dataset with both models, and the better-performing model was selected for each task: Gemini-2.0-flash for synonyms, and GPT-4o for collocations. Evaluation results showed strong performance of the models, with over 80% of the generated distractors rated as appropriate. Common issues included non-existent or rare words and legitimate synonyms in the synonym task, and common collocations or distractors that improperly altered collocational structure in the collocation task. Additional filtering of the data was required to ensure game readiness. Further plans include using LLMs for the production of data for other games, as well as using LLM in the preparation of lexicographic data in the DDDS.

Keywords: language game; LLM; synonym; distractor; collocation; dictionary database

References

Alhazmi, E., Sheng, Q. Z., Zhang, W. E., Zaib, M. & Alhazmi, A. (2024). Distractor Generation in Multiple-Choice Tasks: A Survey of Methods, Datasets, and Evaluation. The 2024 Conference on Empirical Methods in Natural Language Processing. Available at: https://doi.org/10.48550/arXiv.2402.01512.

Arhar Holdt, Š., Čibej, J., Dobrovoljc, K., Gantar, A., Gorjanc, V., Klemenc, B., Kosem, I., Krek, S., Laskowski, C. & Robnik Šikonja, M. Thesaurus of Modern Slovene: By the Community for the Community. In J. Čibej, V. Gorjanc, I. Kosem, S. Krek (eds.). *Proceedings of the XVIII EURALEX International Congress:*

Lexicography in Global Contexts. Ljubljana: Znanstvena založba Filozofske fakultete, pp. 401–410. Available at: https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/ 118/211/3000-1.

Arhar Holdt, Š., Gantar, P., Kosem, I., Pori, E., Robnik Šikonja, M. & Krek, S. (2023). Thesaurus of Modern Slovene 2.0. In *Proceedings of eLex 2023: Electronic Lexicography in the 21st Century*, Brno, pp. 366–381. Available at: https://elex.link/elex2023/wp-content/uploads/82.pdf.

Arhar Holdt, Š., Gapsa, M., Gantar, P. & Kosem, I. (forthcoming). Potencial ChatGPT-ja pri razvoju Slovarja sopomenk sodobne slovenščine. *Contributions to Contemporary History*.

Arhar Holdt, Š. & Kosem, I. (forthcoming). CJVT Igre: New Word Games Based on the Digital Dictionary Database of Slovene. Proceedings of eLex 2025.

Arhar Holdt, Š., Logar, N., Pori, E., Kosem, I. (2021). Game of words: play the game, clean the database (2021). In Z. Gavriilidou, L. Mitits & S. Kiosses (eds.) *Lexicography for inclusion: EURALEX XIX: Congress of the European Association for Lexicography*: 7-9 September 2021, virtual: proceedings book. Vol. 2. Komotini: Democritus University of Thrace, pp. 41–49. Available at: https://euralex.org/publications/game-of-words-play-the-game-clean-the-database/.

Barrett, G. (2023). 'Defin-O-Bots: Challenging A.I. to Create Usable Dictionary Content.' Paper presented at the 24th Biennial Conference of the Dictionary Society of North America. Boulder, CO, USA, 31 May – 3 June 2023.

Collins, J. (2006). Writing multiple-choice questions for continuing medical educa-tion activities and self-assessment modules. *Radiographics*, 26, pp. 543-551. Available at: http://pubs.rsna.org/doi/pdf/10.1148/rg.262055145.

Davies, M. (2025a). Comparing the predictions of Large Language Models to actual corpus data. (White papers). English-Corpora.org. Available at: https://www.english-corpora.org/ai-llms/.

Davies, M. (2025b). Corpora and LLMs: comparing data on word frequency. (White paper). English-Corpora.org. Available at: https://www.english-corpora.org/ai-llms/word-frequency.pdf.

Davies, M. (2025c). Corpora and LLMs: comparing data on phrase frequency. (White paper). English-Corpora.org. Available at: https://www.english-corpora.org/ai-llms/phrase-frequency.pdf.

Davies, M. (2025d). Corpora and LLMs: comparing collocates data. (White paper). English-Corpora.org. Available at: https://www.english-corpora.org/ai-llms/collocates.pdf.

De Schryver, G.-M. (2023). Generative AI and Lexicography: The Current State of the Art Using ChatGPT. *International Journal of Lexicography*, 36, pp. 355–387.

De Schryver, G.-M. & Joffe, D. (2023). 'The End of Lexicography, Welcome to the Machine: On How ChatGPT Can Already Take over All of the Dictionary Maker's Tasks.' Paper presented at the 20th CODH Seminar. Center for Open Data in the Humanities, Research Organization of Information and Systems, National

Institute of Informatics, Tokyo, Japan, 27 February 2023. Accessed at: https://youtu.be/watch?v=mEorw0yefAs.

De Schryver, G.-M. (2024). The Future of the Dictionary. In E. Finegan & M. Adams (eds.) *The Cambridge Handbook of the Dictionary*. Cambridge: Cambridge University Press.

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

Gantar, P., Kosem, I. & Krek, S. (2016). Discovering automated lexicography: the case of Slovene lexical database. *International Journal of Lexicography*, 29(2), pp. 200–225. Available at: https://doi.org/10.1093/ijl/ecw014.

Gierl, M.J., Bulut, O., Guo, Q. & Zhang, X. (2017). Developing, Analyzing, and Using Distractors for Multiple-Choice Tests in Education: A Comprehensive Review. *Review of Educational Research* 87(6), pp. 1082–1116. https://doi.org/10.3102/0034654317726529.

Haladyna, T.M. & Rodriguez, M.C. (2013). *Developing and validating test items*. New York, NY: Routledge.

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 (eds.) *Proceedings of the eLex 2023 Conference: Electronic Lexicography in the 21st Century*. Brno: Lexical Computing, pp. 508–523. Accessed at: https://www.youtube.com/watch?v=8e52vvDpdfQ.

Kosem, I., Arhar Holdt, Š., Gantar, P. & Krek, S. (2023). Collocations Dictionary of Modern Slovene 2.0. In M. Medveď et al. (ed.) *eLex 2023: electronic lexicography in the 21st century (eLex 2023): proceedings of the eLex 2023 conference*, 27–29 June 2023. Brno: Lexical Computing CZ, pp. 491–507. Available at: https://elex.link/elex2023/wp-content/uploads/100.pdf.

Kosem, I., Gantar, P., Arhar Holdt, Š., Gapsa, M., Zgaga, K. & Krek, S. (2024). AI in Lexicography at the University of Ljubljana: case studies. In S. Krek (ed.) *Book of abstracts of the workshop Large Language Models and Lexicography*. 8. October 2024, Cavtat, Croatia, pp. 29–32.

Kosem, I., Krek, S. & Gantar, P. (2021). Semantic data should no longer exist in isolation: the Digital Dictionary Database of Slovenian. In Z. Gavriilidou, L. Mitits & S. Kiosses (eds.) *EURALEX XIX, Congress of the European Association for Lexicography, Lexicography for inclusion: 7–9 September 2021, virtual: abstracts book*. Komotini: Democritus University of Thrace, pp. 81–83. Available at: https://euralex2020.gr/wp-content/uploads/2021/09/EURALEX2020_BookOfAbstracts-Preview-1.pdf.

Kosem, I., Zingano Kuhn, T., Arhar Holdt, Š., Koppel, K., Tiberius, C., Zviel-Girshin, R., Waszink, V. & Zgaga, K. (2024). Can AI assist in Selecting Dictionary Examples? A Case Study in Four Languages. In K. Štrkalj Despot, A. Ostroški Anić & I. Brač (eds.) *Lexicography and Semantics. Book of abstracts of the XXI*

EURALEX International Congress. Institut za hrvatski jezik, pp. 128–130.

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.) *LREC 2020: Twelfth International Conference on Language Resources and Evaluation*: May 11-16, 2020, Marseille, France. Paris: ELRA - European Language Resources Association, pp. 3340–3345. Available at: http://www.lrec-conf.org/proceedings/lrec2020/LREC-2020.pdf.

Krek, S., Laskowski, C. & Robnik-Šikonja, M. (2017). From translation equivalents to synonyms: creation of a Slovene thesaurus using word co-occurrence network analysis. In I. Kosem et al. (eds.) *Proceedings of eLex 2017: Lexicography from Scratch*, 19-21 September 2017, Leiden, Netherlands. Available at: https://elex.link/elex2017/wp-content/uploads/2017/09/paper05.pdf.

Lew, R. (2023). ChatGPT as a COBUILD lexicographer. *Humanities and Social Sciences Communications* 10, no. 704. Available at: https://doi.org/10.1057/s41599-023-02119-6.

McKean, E. & Fitzgerald, W. (2023). The ROI of AI in Lexicography. *Proceedings of the 16th International Conference of the Asian Association for Lexicography: "Lexicography, Artificial Intelligence, and Dictionary Users"*. Seoul: Yonsei University, pp. 10–20.

Mitkov, R., Ha, L.A., & Karamanis, N. (2006). A computer-aided environment forgenerating multiple-choice test items. Natural Language Engineering,12, pp. 177–194. Available at: https://doi.org/10.1017/S1351324906004177.

Nichols, W. (2023). Invisible Lexicographers, AI, and the Future of the Dictionary. Paper presented at the eLex 2023 Conference: Electronic Lexicography in the 21st Century. Brno, Czech Republic, 27-29 June 2023. Accessed at: https://www.youtube.com/watch?v=xYpwftj_QQI.

Ratcliff, J.W. & Metzener, D. (1988). Pattern Matching: The Gestalt Approach. *Dr. Dobb's Journal*, pp. 46.

Špica, D. & Perak, B. (2024). Enhancing Japanese Lexical Networks Using Large Language Models – Extracting Synonyms and Antonyms with GPT-4o. In K. Štrkalj Despot, A. Ostroški Anić & I. Brač (eds.) *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress*. Institut za hrvatski jezik, pp. 283–303.

Thissen, D., Steinberg, L. & Fitzpatrick, A.R. (1989). Multiple-choice models: Thedistractors are also part of the item. *Journal of Educational Measurement*, 26, pp. 161–176. Available at: https://doi.org/10.1111/j.1745-3984.1989.tb00326.x.

Tiberius, C., Heylen, K., de Does, J., Vanroy, B., Vandeghinste, V. & van Doeselaar, J. (2024). LLMs and Evidence-Based Lexicography: Pilot studies at INT. In S. Krek (ed.) *Book of abstracts of the workshop Large Language Models and Lexicography*. 8. October 2024, Cavtat, Croatia, pp. 49–52.

Tran, H. T. H., Podpečan, V., Jemec Tomazin, M. & Pollak, S. (2023). Definition Extraction for Slovene: Patterns, Transformer Classifiers and ChatGPT. In M. Medveď, M. Měchura, C. Tiberius, I. Kosem, J. Kallas & M. Jakubíček (eds.)

Proceedings of the eLex 2023 Conference: Electronic Lexicography in the 21st Century. Brno: Lexical Computing, pp. 19–38. Accessed at: https://www.youtube.com/watch?v=rQC3Rz04b20.

Tuulik, M., Risberg, L., Koppel, K., Aedmaa, E., Prangel, E., Zupping, S., Vainik, E. & Langemets, M. (2024). Who are Better at Semantics – Experienced Lexicographers or LLMs? In K. Štrkalj Despot, A. Ostroški Anić & I. Brač (eds.) *Lexicography and Semantics. Book of abstracts of the XXI EURALEX International Congress*. Institut za hrvatski jezik, pp. 219–221.

Resources:

OpenAI. (2023). *ChatGPT* (March 2025 version, ChatGPT-4o) [Large language model]. Accessed at: https://chat.openai.com/chat.

Google Gemini. (version 2.0-flash; API; February 2025) Accessed at: https://gemini.google.com/.

DMLEX on Wikibase: Legacy dictionaries as collaboratively editable dataset

Simon Krek[1,2], Primož Ponikvar[3], Andraž Repar[2], Iztok Kosem[1,2], David Lindemann[4]

[1] University of Ljubljana (Faculty of Arts Faculty of Computer and Information Science), Ljubljana, Slovenia

[2] Jožef Stefan Institute, Ljubljana, Slovenia

[3] Centre for Language Resources and Technologies at the University of Ljubljana, Ljubljana, Slovenia

[4] EHU University of the Basque Country, Vitoria-Gasteiz, Spain

E-mail: simon.krek@ijs.si, pponikvar@yahoo.com, andraz.repar@ijs.si, iztok.kosem@ijs.si, david.lindemann@ehu.eus

Abstract

This paper presents an experimental workflow for converting legacy digitized dictionaries into the DMLex standard and subsequently importing them into a Wikibase instance. DMLex, a serialization-independent model developed by the OASIS LEXIDMA Technical Committee, aims to provide a universal and modular representation of lexicographic data. The study tested whether dictionaries from heterogeneous sources—originally encoded in internal XML formats—could be reliably transformed into DMLex-compliant representations and repurposed for collaborative editing and enrichment on a structured linked data platform. The transformation was achieved through a combination of rule-based scripts, manual refinement, and large language model assistance. While DMLex proved adaptable to a wide range of lexical phenomena, several limitations became apparent during the Wikibase integration phase. These findings suggest that practical deployment of DMLex benefits from clearer conventions and validation strategies when applied beyond theoretical modeling. The results confirm DMLex's potential for future-proof dictionary modeling, while also highlighting areas where further specification and community consensus are needed to support its application in digital infrastructures and collaborative environments.

Keywords: legacy dictionaries; conversion; standardization; semantic web; linked data

References

Almeida, B., Costa, R., Salgado, A., Ramos, M., Romary, L., Khan, F., Carvalho, S., Khemakhem, M., Silva, R. & Tasovac, T. (2022). Modelling usage information in

a legacy dictionary: from TEI Lex-0 to Ontolex-Lemon. In *Workshop on Computational Methods in the Humanities 2022 (COMHUM 2022)*, volume 3602. CEUR Workshop proceedings.

Belyaev, O., Khomchenkova, I., Sinitsyna, J. & Dyachkov, V. (2021). Digitizing print dictionaries using TEI: The Abaev Dictionary Project. In *Proceedings of the Seventh International Workshop on Computational Linguistics of Uralic Languages*, pp. 57–64.

Biffi, M., Sassolini, E., Monachini, M., Montemagni, S. et al. (2019). Converting and structuring a digital historical dictionary of Italian: a case study. In *Electronic lexicography in the 21st century: smart lexicography. Proceedings of the eLex 2019 conference (1-3 October 2019, Sintra, Portugal)*. Lexical Computing CZ, pp. 603–621.

Francopoulo, G. (2013). *LMF lexical markup framework*. John Wiley & Sons.

Kosem, I. & Lindemann, D. (2021). New developments in Elexifinder, a discovery portal for lexicographic literature. In Z. Gavriilidou, L. Mitits & S. Kiosses (eds.) *Lexicography for Inclusion: Proceedings of the 19th EURALEX International Congress, 7-11 September 2021, Alexandroupolis, Vol. 2.* Alexandroupolis: Democritus University of Thrace, pp. 759–766. Available at: https://euralex2020.gr/proceedings-volume-2/.

Lindemann, D. (2025). Ontolex-Lemon in Wikidata and other Wikibase instances. In *Fifth Ontolex Workshop, September 9, 2025*. Naples: Zenodo. Available at: https://doi.org/10.5281/zenodo.15471514.

Lindemann, D., Ahmadi, S., Khan, A.F., Mambrini, F., Iurescia, F. & Passarotti, M.C. (2023). When OntoLex Meets Wikibase: Remodeling Use Cases. *CEUR Workshop proceedings*, 2773. Available at: https://ceur-ws.org/Vol-3640/ paper14.pdf.

Maxwell, M. & Bills, A. (2017). Endangered data for endangered languages: Digitizing print dictionaries. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*.

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: Lexicography from scratch. Proceedings of eLex 2017*. Brno: Lexical Computing CZ s.r.o., pp. 587–597. Available at: https://elex.link/elex2017/ wp-content/uploads/2017/09/paper36.pdf.

Romary, L. & Lopez, P. (2015). Grobid-information extraction from scientific publications. *ERCIM News*, 100.

Salgado, A. (2018). From Legacy Formats and Databases to TEI: Converting the Academy of Sciences Portuguese Dictionary to TEI Lex-0. DigiLex. Available at: https://doi.org/10.58079/nmo1. (20 July 2025)

Tasovac, T., Romary, L., Banski, P., Bowers, J., de Does, J., Depuydt, K., Erjavec, T., Geyken, A., Herold, A., Hildenbrandt, V., Khemakhem, M., Lehečka, B., Petrović, S., Salgado, A. & Witt, A. (2018). TEI Lex-0: A Baseline Encoding for Lexicographic Data. DARIAH Working Group on Lexical Resources. Available at:

https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html.

Tiberius, C., Krek, S., Depuydt, K., Gantar, P., Kallas, J., Kosem, I. & Rundell, M. (2021). Towards the ELEXIS data model: defining a common vocabulary for lexicographic resources. *Electronic lexicography in the 21st century (eLex 2021) Post-editing lexicography*, 91.

Neology in Practice: Lexicographic and Terminological Approaches to Lexical Innovation

Jelena Kallas[1], Kristina Koppel[1], Kris Heylen[2], Ilan Kernerman[3], Ana Ostroški Anić[4], Federica Vezzani[5], Špela Arhar Holdt[6]

[1] Institute of the Estonian Language, Tallinn, Estonia
[2] Instituut voor de Nederlandse Taal, Leiden, the Netherlands
[3] Lexicala by K Dictionaries, Tel Aviv, Israel
[4] Institute for the Croatian Language, Zagreb, Croatia
[5] Università degli Studi di Padova, Padua, Italy
[6] University of Ljubljana, Ljubljana, Slovenia
E-mail: jelena.kallas@eki.ee, kristina.koppel@eki.ee, kris.heylen@ivdnt.org, ilan@lexicala.com, aostrosk@ihjj.hr, federica.vezzani@unipd.it, spela.arharholdt@ff.uni-lj.si

Abstract

The COST Action 'European Network on Lexical Innovation' (ENEOLI) has conducted a comprehensive survey in October-November 2024 regarding the methods, practices, tools, and resources used in the study and documentation of lexical innovations, including neologisms and novel senses. The 249 respondents from 50 countries represented linguists, lexicographers, terminologists, translators, software developers, and educators. Respondents could indicate more than one field of expertise, and 169 noted theirs as linguistics (70%), 107 lexicography (44%) and 105 terminology (43%). In this paper, we focus on the responses of those indicating their field of expertise as lexicography and/or terminology, and we analyzed their approaches to the identification and documentation of neologisms, the composition of project teams and the use of corpora and digital tools. Special attention is given to training pathways and professional needs, offering insights into the evolving skills required in the field of lexical innovation.

Keywords: neologisms; neology; lexical innovation; lexicography; terminology; survey

References

Anthony, L. (2013). Developing AntConc for a new generation of corpus linguists. In *Proceedings of the Corpus Linguistics Conference (CL 2013)*, pp. 14–16. Available at: https://laurenceanthony.net/research/20130722_26_cl_2013/cl_2013_paper_final.pdf.

Borin, L., Forsberg, M. & Roxendal, J. (2012). Korp – the corpus infrastructure of Språkbanken. In *Proceedings of the Eighth International Conference on Language*

Resources and Evaluation (LREC'12), *Istanbul, Turkey*. European Language Resources Association (ELRA), pp. 474–478.

Cartier, E. (2017). Neoveille, a Web Platform for Neologism Tracking. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Valencia, Spain*. Association for Computational Linguistics, pp. 95–98.

Cassotti, P., Siciliani, L., de Gemmis, M., Semeraro, G. & Basile, P. (2023). Xl-lexeme: Wic pretrained model for cross-lingual lexical semantic change. In *Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, pp. 1577–1583.

Di Nunzio, G.M. & Vezzani, F. (2025). FAIRterm 2.0: Towards FAIR terminologies resources for EOSC. *IEEE International Conference on Cyber Humanities (IEEE-CH), IEEE Explore*. IEEE.

François, M. (2024). Exploitation d'un corpus oral en anglais dans le domaine de la modélisation climatique: contribution à la description d'une néologie «première». *Approches épistémologiques de la terminologie - Enjeux actuels*, Réseau LTT - Lexicologie, Terminologie, Traduction, Oct 2024, Paris, France.

Freixa, J. (2022a). 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*, Peter Lang, pp. 15–37.

Freixa, J. (2022b). 'Garbell: L'avaluador Automàtic de Neologismes Catalans (Garbell: The Automatic Catalans Neologisms Analyzer)'. *Terminàlia*, 26, pp. 7–16.

Kerremans, D., Stegmayr, S. & Schmid, H.-J. (2012). The NeoCrawler: Identifying and retrieving neologisms from the Internet and monitoring ongoing change. In K. Allan & J.A. Robinson (eds.) *Current methods in historical semantics*. De Gruyter Mouton, pp. 59–96.

Kilgarriff, A., Rychlý, P., Smrž, P. & Tugwell, D. (2004). The Sketch Engine. In G. Williams & S. Vessier (eds.) *Proceedings of the 11th EURALEX International Congress*. Université de Bretagne-Sud, pp. 105–115.

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M. & Rychlý, P. (2008). GDEX: Automatically finding good dictionary examples in a corpus. In *Proceedings of the 13th EURALEX International Congress*. Spain, July 2008, pp. 425–432.

Klosa-Kückelhaus, A. & Rüdiger, J. O. (2023). Introducing NeoRate. *Lexicography*, 10(2), pp. 117–137.

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.) *Proceedings of the eLex 2019 conference. 1-3 October 2019, Sintra, Portugal*. Brno: Lexical Computing CZ, s.r.o., pp. 434−452.

Kranebitter, K. & Ralli, N. (2022). Quanto può influire l'utente nello sviluppo di uno strumento terminologico? L'esperienza di bistro. *Risorse e strumenti per*

l'elaborazione e la diffusione della terminologia in Italia. Eurac Research, pp. 102−116.

Nimb, S., Sørensen, N.H. & Lorentzen, H. (2020). Updating the dictionary: Semantic change identification based on change in bigrams over time. *Slovenščina 2.0*: *empirične, aplikativne in interdisciplinarne raziskave*, 8(2), pp. 112–138.

Salgado, A., Simões, A., Iriarte, A., Vieira, R., Ferreira, M., Carmo, R., Pinheiro, C. (2023). Dicionário da Língua Portuguesa: a new lexicographic resource of Academia das Ciências de Lisboa. *Electronic lexicography in the 21st century (eLex 2023): Invisible Lexicography. Proceedings of the eLex 2023 conference. Brno, 27–29 June 2023.* Brno: Lexical Computing CZ s.r.o., pp. 72–75.

Schlechtweg, D., Virk, S.M., Sander, P., Sköldberg, E., Linke, L.T., Zhang, T., Tahmasebi, N., Kuhn, J. & Walde, S.S.I. (2023). The DURel Annotation Tool: Human and Computational Measurement of Semantic Proximity, Sense Clusters and Semantic Change. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, St. Julians, Malta. Association for Computational Linguistics*, pp. 137–149.

Sköldberg, E., Virk, S.M., Sander, P., Hengchen, S. & Schlechtweg, D. (2024). Semantic Variation in Swedish Using Computational Models of Semantic Proximity – Results From Lexicographical Experiments. In K.Š. Despot, A. Ostroški Anić & I. Brač (eds.) *Proceedings of the XXI EURALEX International Congress: Lexicography and Semantics.* Cavtat (Croatia): Institute for the Croatian Language, 2024, pp. 169–182.

Sørensen, N.H. & Nimb, S. (2018). Word2Dict – Lemma Selection and Dictionary Editing Assisted by Word Embeddings. In J. Čibej, V. Gorjanc, I. Kosem & S. Krek (eds.) *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts.* Ljubljana University Press, Faculty of Arts, pp. 819–826.

Storjohann, P. (2024). IDS-Neo 2020+: A Novel Resource for New German Words in Use. *International Journal of Lexicography,* 37(4), pp. 389–403.

Suonuuti, H. (1997). *Guide to terminology.* Helsinki: Tekniikan Sanastokeskus.

Tavast, A., Langemets, M., Kallas, J. & Koppel, K. (2018, July). Unified data modelling for presenting lexical data: The case of EKILEX. In J. Čibej, V. Gorjanc, I. Kosem & Simon Krek (eds.) *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts.* Ljubljana, pp. 749–761.

Tiberius, C., Kallas, J., Koeva, S. & Langemets, M. (2024). A lexicographic practice map of Europe. *International Journal of Lexicography*, 37(1), pp. 1–28.

Trap-Jensen, L. (2016). Lexicography at the Society for Danish Language and Literature. *Kernerman Dictionary News*, 16.

Vezzani, F., Di Nunzio, G. M., & Henrot, G. (2018). Trimed: A multilingual terminological database. *LREC 2018-11th International Conference on Language Resources and Evaluation*, pp. 4367–4371.

Vezzani, F. (2021). La ressource FAIRterm: entre pratique pédagogique et

professionnalisation en traduction spécialisée. *Synergies Italie*, 17, pp. 51–64.

Waszink, V. (2020). Neologisms in an online portal: The Dutch Neologismenwoordenboek (NW). *Dictionaries: Journal of the Dictionary Society of North America 41, no. 1 (2020)*, pp. 27–44.

Exploring the constructicographic potential of lexicographic data and language models: The case of the Estonian Nominal Quantifier Construction

Heete Sahkai[1], Geda Paulsen[1,2], Ene Vainik[1], Jelena Kallas[1],

Ahto Kiil[3], Katrin Tsepelina[1,3], Kertu Saul[1,3], Arvi Tavast[1]

[1] Institute of the Estonian Language, Roosikrantsi 6, Tallinn 10119, Estonia
[2] Uppsala University, Thunbergsvägen 3 L, Uppsala 75126, Sweden
[3] University of Tartu, Ülikooli 18, Tartu 50090, Estonia
E-mail: heete.sahkai@eki.ee, geda.paulsen@eki.ee, ene.vainik@eki.ee, jelena.kallas@eki.ee, ahto.kiil@gmail.com, katrin,tsepelina@eki.ee, kertu.saul@eki.ee, arvi.tavast@eki.ee

Abstract

Constructicography, or the description of grammatical constructions in a lexicographic format, is an emerging field currently in the stage of developing and automating methods for treating large numbers of (semi-)schematic constructions. This study explores how existing lexicographic data and language models can be used to facilitate the constructicographic workflow. Our results suggest that (1) collocations and semantic relations represented in a lexicographic database can be used to identify the collexemes of constructions, that is, the lexemes occurring in the open slot(s) of schematic constructions, (2) BERT-based language models can be trained to identify instances of constructions in corpora, using collocations as the starting point to create appropriate training data, and (3) commercial large language models can be prompted to identify constructional instances, using a small number of examples. The identification of the collexemes and corpus instances of constructions provide several pieces of information that can be represented in constructicon entries: the meaning, form, frequency and productivity of constructions, the frequency and association strength of particular collexemes, the CEFR-level of the construction, etc.

Keywords: Constructicography; BERT-based models; Large Language Models; Lexicography; Collostructional Analysis; Estonian

References

Baayen, H. (2009). Corpus linguistics in morphology: Morphological productivity. *Corpus Linguistics: An International Handbook*, pp. 899–919. Available at: https://doi.org/10.1515/9783110213881.2.899.

Barteld, F. & Ziem, A. (2020). Construction mining: Identifying construction candidates for the German constructicon. In *Belgian Journal of Linguistics*, 34, pp. 5–16. Available at: https://doi.org/10.1075/bjl.00030.bar.

Bonial, C. & Tayyar Madabushi, H. (2024). Constructing understanding: On the constructional information encoded in large language models. In *Language Resources and Evaluation*. Available at: https://doi.org/10.1007/s10579-024-09799-9.

Borin, L. & Lyngfelt, B. (2025). Framenets and constructiCons. *The Cambridge Handbook of Construction Grammar*. Cambridge University Press. Available at: https://www.academia.edu/95301779/Framenets_and_constructiCons.

Bäckström, L., Borin, L., Forsberg, M., Lyngfelt, B., Prentice, J. & Sköldberg, E. (2013). Automatic identification of construction candidates for a Swedish constructicon. In *Proceedings of the workshop on lexical semantic resources for NLP at NODALIDA 2013. NEALT Proceedings Series 19 / Linköping Electronic Conference Proceedings 88*, pp. 2–11.

Dunn, J. (2017). Computational learning of construction grammars. In *Language and Cognition: An Interdisciplinary Journal of Language and Cognitive Science*, 9(2), pp. 254–292. Available at: https://doi.org/10.1017/langcog.2016.7.

Dunn, J. (2023). *Exploring the Constructicon: Linguistic Analysis of a Computational CxG.* Available at: https://doi.org/10.48550/arxiv.2301.12642.

Fillmore, C.J. (2006, September 3). *The articulation of lexicon and constructicon* [Plenary lecture]. Fourth International Conference on Construction Grammar (ICCG4), University of Tokyo, Japan.

Fillmore, Charles J. (2008). Border Conflicts: FrameNet Meets Construction Grammar. In *Proceedings of the XIII EURALEX International Congress.* Barcelona: Universitat Pompeu Fabra, pp. 49–68.

Fillmore, C.J., Kay, P., O'Connor, M.C. (1988). Regularity and idiomaticity in grammatical constructions. In *Language* 64, pp. 501–538.

Forsberg, M., Johansson, R., Bäckström, L., Borin, L., Lyngfelt, B., Olofsson, J. & Prentice, J. (2014). From construction candidates to constructicon entries: An experiment using semi-automatic methods for identifying constructions in corpora. In *Constructions and Frames*, 6(1), pp. 114–135. Available at: https://doi.org/10.1075/cf.6.1.07for.

Gries, S.T. (2022). *Coll.analysis 4.0. A script for R to compute perform collostructional analyses*. Available at: https://www.stgries.info/teaching/groningen/index.html.

Goldberg, A.E. (1995). *Constructions: A Construction Grammar Approach to Argument Structure*. University of Chicago Press.

Herbst, T. & Hoffmann, T. (2024). *A Construction Grammar of the English Language: CASA – a Constructionist Approach to Syntactic Analysis*. John Benjamins Publishing Company. Available at: https://doi.org/10.1075/clip.5.

Janda, L.A., Endresen, A., Zhukova, V., Mordashova, D., & Rakhilina, E. (2020). How to build a constructicon in five years. The Russian example. In *Belgian Journal of Linguistics*, 34(1), pp. 161–173. Available at: https://doi.org/10.1075/

bjl.00043.jan.

Kallas, J. (2013). *Eesti keele sisusõnade süntagmaatilised suhted korpus- ja õppeleksikograafias* [Tallinn University. Dissertations on humanities]. Tallinna Ülikool.

Kallas, J., Kilgarriff, A., Koppel, K., Kudritski, E., Langemets, M., Michelfeit, J., Tuulik, M. & Viks, Ü. (2015). Automatic generation of the Estonian Collocations Dictionary database. In *Electronic lexicography in the 21st century: Linking lexical data in the digital age. Proceedings of the eLex 2015 conference, 11–13 August 2015, Herstmonceux Castle, United Kingdom*. Trojina, Institute for Applied Slovene Studies / Lexical Computing Ltd., pp. 49–68.

Kay, P. & Fillmore, C.J. (1999) Grammatical constructions and linguistic generalizations: The What's X doing Y? construction. In *Language* 75, pp. 1–33.

Kilgarriff, A., Rychly, P., Smrž, P. & Tugwell, D. (2004). The Sketch Engine. In *Proceedings of the XI Euralex International Congress*. Lorient: Université de Bretagne Sud, pp. 105–116.

Kilgarriff, A., Rychlý, P., Jakubicek, M., Kovář, V., Baisa, V. & Kocincová, L. (2014). Extrinsic Corpus Evaluation with a Collocation Dictionary Task. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation.* Reykjavik, Iceland, pp. 454–552.

Koppel, K. & Kallas, J. (2022). Eesti keele ühendkorpuste sari 2013–2021: Mahukaim eestikeelsete digitekstide kogu. In *Eesti Rakenduslingvistika Ühingu aastaraamat*, *18*, pp. 207–228. Available at: https://doi.org/10.5128/ERYa18.12.

Koppel, K., Tavast, A., Langemets, M. & Kallas, J. (2019). Aggregating dictionaries into the language portal Sõnaveeb: Issues with and without a solution. In *Proceedings of the eLex 2019 conference.* Lexical Computing CZ, s.r.o., pp. 434−452.

Koptjevskaja-Tamm, M. (2001). Partitive and pseudo-partitive nominal constructions in the Circum-Baltic languages: "A piece of the cake" and "a cup of tea". In Ö. Dahl & M. Koptjevskaja-Tamm (eds.) *Circum-Baltic Languages: Volume 2: Grammar and Typology.* John Benjamins Publishing Company, pp. 523–568. Available at: https://doi.org/10.1075/slcs.55.11kop.

Lee-Goldman, R. & Petruck, M.R.L. (2018). The FrameNet constructicon in action. *Constructicography: Constructicon development across languages.* John Benjamins Publishing Company, pp. 19–40. Available at: https://doi.org/10.1075/cal.22.02lee.

Lyngfelt, B., Bäckström, L., Borin, L., Ehrlemark, A. & Rydstedt, R. (2018). Constructicography at work. Theory meets practice in the Swedish constructicon. *Constructicography: Constructicon development across languages.* John Benjamins Publishing Company, pp. 41–106. Available at: https://doi.org/10.1075/cal.22.01lyn.

Metslang, H. 2017. Kvantorifraas. – M. Erelt & H. Metslang (eds.) *Eesti keele süntaks. Eesti keele varamu, 3*. Tartu: Tartu Ülikooli Kirjastus, pp. 463–478.

Patel, M., Garibyan, A., Winckel, E. & Evert, S. (2023). A reference constructicon as

a database. In *Yearbook of the German Cognitive Linguistics Association*, 11(1), pp. 175–202. Available at: https://doi.org/10.1515/gcla-2023-0009.

Perek, F., & Patten, A. L. (2019). Towards an English Constructicon using patterns and frames. In *International Journal of Corpus Linguistics*, 24(3), pp. 354–384. Available at: https://doi.org/10.1075/ijcl.00016.per.

Pilvik, M.-L., Lindström, L., Plado, H. & Simmul, C.E. (2025). Nimisõnafraasi ja hulgafraasi piirimail: „osa", „enamik" ja „enamus" hulgasõnadena. In *Eesti Rakenduslingvistika Ühingu Aastaraamat*, 21, pp. 237–261. Available at: https://doi.org/10.5128/ERYa21.13.

Risberg, L., Tuulik, M., Langemets, M., Koppel, K., Vainik, E., Prangel, E. & Aedmaa, E. (in press). Keelekorpus kui leksikograafi abiline kõnekeelsuse tuvastamisel [Using corpus data to support lexicographers in identifying informal language]. Keel ja Kirjandus, 7.

Sass, B. (2023). From a dictionary towards the Hungarian Constructicon. In *Electronic Lexicography in the 21st Century (eLex 2023): Invisible Lexicography. Proceedings of the eLex 2023 Conference. Brno, 27–29 June 2023.* Brno: Lexical Computing CZ s.r.o., pp. 534–544.

Shibuya, Y. & Jensen, K.E. (2015). Mining for constructions in texts using N-gram and network analysis. In *Globe: A Journal of Language, Culture and Communication,* 2, pp. 23–54.

Sidorov, G. (2019). *Syntactic n-grams in Computational Linguistics.* Springer International Publishing. Available at: https://doi.org/10.1007/978-3-030-14771-6.

Stefanowitsch, A. & Gries, S.T. (2003). Collostructions: Investigating the interaction of words and constructions. In *International Journal of Corpus Linguistics*, 8(2), pp. 209–243. Available at: https://doi.org/10.1075/ijcl.8.2.03ste.

Ziem, A. & Feldmüller, T. (2023). Dimensions of constructional meanings in the German Constructicon: Why collo-profiles matter. In *Yearbook of the German Cognitive Linguistics Association*, 11(1), pp. 203–226. Available at: https://doi.org/10.1515/gcla-2023-0010.

Ziem, A., Flick, J. & Sandkühler, P. (2019). The German Constructicon Project: Framework, methodology, resources. In *Lexicographica*, 35, pp. 15–40. Available at: https://doi.org/10.1515/lex-2019-0003.

Tanvir, H., Kittask, C., Eiche, S. & Sirts, K. (2021). EstBERT: A Pretrained Language-Specific BERT for Estonian. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pp. 11–19.

Tavast, A., Koppel, K., Langemets, M. & Kallas, J. (2020). Towards the Superdictionary: Layers, Tools and Unidirectional Meaning Relations. In *Proceedings of XIX EURALEX Congress: Lexicography for Inclusion, Vol. I.* Alexandroupolis, Greece: Democritus University of Thrace, pp. 215—223.

Tavast, A., Langemets, M., Kallas, J. & Koppel, K. (2018). Unified Data Modelling for Presenting Lexical Data: The Case of EKILEX. In *Proceedings of the XVIII EURALEX International Congress: EURALEX: Lexicography in Global Contexts,*

Ljubljana, 17-21 July 2018. Ljubljana University Press, Faculty of Arts, pp. 749−761.

Ulčar, M. & Robnik-Šikonja, M. (2021). *Training dataset and dictionary sizes matter in BERT models: The case of Baltic languages.* Available at: https://doi.org/10.48550/arXiv.2112.10553; Version 1.

Vainik, E., Paulsen, G., Sahkai, H., Kallas, J., Tavast, A. & Koppel, K. (2024). From a Dictionary to a Constructicon – Putting the Basics on the Map. In *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language, pp. 209–216.

Wible, D. & Tsao, N.-L. (2010). StringNet as a Computational Resource for Discovering and Investigating Linguistic Constructions. In *Proceedings of the NAACL HLT Workshop on Extracting and Using Constructions in Computational Linguistics.* Los Angeles, California, Association for Computational Linguistics, pp. 25–31.

Corpora and datasets

Balanced Corpus of Estonian. Accessed at: https://www.cl.ut.ee/korpused/grammatikakorpus/. (1 October 2025)

Estonian as a Second Language Coursebook Sentences Corpus 2021. Accessed at: https://doi.org/10.15155/3-00-0000-0000-0000-0885AL. (1 October 2025)

Estonian as a Second Language School Coursebook Sentences Corpus. 2021 Accessed at: https://doi.org/10.15155/3-00-0000-0000-0000-0888DL. (1 October 2025)

Estonian National Corpus 2017. Accessed at: https://doi.org/10.15155/3-00-0000-0000-0000-071E7L. (1 October 2025)

Estonian National Corpus 2021. Accessed at: https://doi.org/10.15155/3-00-0000-0000-0000-08D17L. (1 October 2025).

Estonian National Corpus 2023. Accessed at: https://doi.org/10.15155/3-00-0000-0000-0000-08C04M. (1 October 2025)

Estonian Reference Corpus. Accessed at: https://www.cl.ut.ee/korpused/segakorpus/index.php?lang=en. (1 October 2025)

Nominal quantifier constructions_ Gold Standard Dataset. Accessed at: https://github.com/keeleinstituut/PRG1978/tree/main/constructions/gold_standards.

Language models

CamemBERT. Accessed at: https://huggingface.co/docs/transformers/model_doc/camembert.

Claude-Sonnet-4. Accessed at: https://claude.ai.

EstBERT. Accessed at: https://huggingface.co/tartuNLP/EstBERT.

Est-RoBERTa. Accessed at: https://huggingface.co/EMBEDDIA/est-roberta.

GPT-4.1. Accessed at: https://openai.com/index/gpt-4-1/.

o3-mini. Accessed at: https://openai.com/index/openai-o3-mini/.

TartuNLP/EstRoBERTa. Accessed at: https://huggingface.co/tartuNLP/EstRoBERTa.

Automatically Updated Corpora of EU National Parliaments with Terminology Extraction in Twenty Languages

Marek Blahuš[1], Ota Mikušek[1,2]

[1] Lexical Computing, Brno, Czech Republic
[2] Faculty of Informatics, Masaryk University, Brno, Czech Republic
E-mail: firstname.lastname@sketchengine.eu

Abstract

We present a collection of monolingual text corpora derived from the steno protocols of 30 parliamentary chambers across 22 EU member states, covering 20 languages. The corpora are continuously and automatically updated, enabling intralingual and cross-lingual analysis of parliamentary discussions. Each chamber's protocols are regularly downloaded, processed, and transformed into a unified prevertical text format. A terminology extraction grammar is available for each language, allowing the identification of terms specific to each parliament by comparing the parliamentary debates with a general-language reference corpus (or a custom subsection of the debates to the whole body of them). The corpora include timestamps, enabling the observation of trending topics across all European national parliaments within a single platform. Corpus quality depends on the availability and format of the source data, which ranges from simple text files, DOCX, HTML, to XML and JSON (With documented APIs). A monitoring system ensures ongoing compatibility with any format changes. Currently, the corpora consist of over 2.8 billion words and are managed in Sketch Engine.

Keywords: spoken corpora; parliamentary debates; multilingual corpora; terminology extraction; diachronic analysis

References

Aker, A., Paramita, M.L. & Gaizauskas, R. (2013). Extracting bilingual terminologies from comparable corpora. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* pp. 402–411.

Blahuš, M., Jakubíček, M., Cukr, M., Kovář, V. & Suchomel, V. (2023). Development of Evidence-Based Grammars for Terminology Extraction in OneClick Terms. *Electronic lexicography in the 21st century. Proceedings of the eLex 2023 conference*, pp. 650–662.

Erjavec, T., Grigorova, V., Ljubešić, N., Ogrodniczuk, M., Osenova, P., Pančur, A., Rudolf, M. & Simov, K. (2020). Multilingual comparable corpora of parliamentary debates ParlaMint 1.0. Available at: http://hdl.handle.net/11356/1345. Slovenian language resource repository CLARIN.SI.

Erjavec, T., Kopp, M., Kuzman Pungeršek, T., Ljubešić, N., Ogrodniczuk, M., Osenova, P., Agirrezabal, M., Agnoloni, T., Aires, J., Albini, M., Alkorta, J., Antiba-Cartazo, I., Arrieta, E., Barcala, M., Bardanca, D., Barkarson, S., Bartolini, R., Battistoni, R., Bel, N., Bonet Ramos, M.d.M., Calzada Pérez, M., Cardoso, A., Çöltekin, Ç., Coole, M., Darg‘ is, R., de Libano, R., Depoorter, G., Diwersy, S., Dodé, R., Fernandez, K., Fernández Rei, E., Frontini, F., Garcia, M., García Díaz, N., García Louzao, P., Gavriilidou, M., Gkoumas, D., Grigorov, I., Grigorova, V., Haltrup Hansen, D., Iruskieta, M., Jarlbrink, J., Jelencsik-Mátyus, K., Jongejan, B., Kahusk, N., Kirnbauer, M., Kryvenko, A., Ligeti-Nagy, N., Luxardo, G., Magariños, C., Magnusson, M., Marchetti, C., Marx, M., Meden, K., Mendes, A., Mochtak, M., Mölder, M., Montemagni, S., Navarretta, C., Nitoń, B., Norén, F.M., Nwadukwe, A., Ojsteršek, M., Pančur, A., Papavassiliou, V., Pereira, R., Pérez Lago, M., Piperidis, S., Pirker, H., Pisani, M., Pol, H.v.d., Prokopidis, P., Quochi, V., Rayson, P., Regueira, X.L., Rii, A., Rudolf, M., Ruisi, M., Rupnik, P., Schopper, D., Simov, K., Sinikallio, L., Skubic, J., Tungland, L.M., Tuominen, J., van Heusden, R., Varga, Z., Vázquez Abuín, M., Venturi, G., Vidal Miguéns, A., Vider, K., Vivel Couso, A., Vladu, A.I., Wissik, T., Yrjänäinen, V., Zevallos, R. & Fišer, D. (2025). Multilingual comparable corpora of parliamentary debates ParlaMint 5.0. Available at: http://hdl.handle.net/11356/2004. Slovenian language resource repository CLARIN.SI.

Erjavec, T., Ogrodniczuk, M., Osenova, P., Ljubešić, N., Simov, K., Pančur, A., Rudolf, M., Kopp, M., Barkarson, S., Steingrímsson, S., Çöltekin, Ç., de Does, J., Depuydt, K., Agnoloni, T., Venturi, G., Pérez, M.C., de Macedo, L.D., Navarretta, C., Luxardo, G., Coole, M., Rayson, P., Morkevičius, V., Krilavičius, T., Dar´gis, R., Ring, O., van Heusden, R., Marx, M. & Fišer, D. (2022). The ParlaMint corpora of parliamentary proceedings. *Language Resources and Evaluation*. Available at: https://doi.org/10.1007/s10579-021-09574-0.

Gojun, A., Heid, U., Weissbach, B., Loth, C. & Mingers, I. (2012). Adapting and evaluating a generic term extraction tool. In *LREC,* pp. 651–656.

IPU, 2014 (2014). Technological Options for Capturing and Reporting Parliamentary Proceedings. Available at: https://www.ipu.org/resources/publications/reference/ 2016-07/technological-options-capturing-and-reporting-parliamentary-proceedings. Document prepared by the United Nations Department of Economic and Social Affairs and the Inter-Parliamentary Union through the Global Centre for ICT in Parliament.

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychl`y, P. & Suchomel, V. (2014). Finding terms in corpora for many languages with the Sketch Engine. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics,* pp. 53–56.

Kilgarriff, A. (2009). Simple maths for keywords. In *Proceedings of the Corpus Linguistics Conference*, volume 6. University of Liverpool Liverpool, pp. 41–55.

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, pp. 7–36.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *The Tenth Machine Translation Summit Proceedings of Conference*. International Association for Machine Translation, pp. 79–86.

Mikušek, O. (2023). Continuous automatic development of European parliamentary corpora [online]. Available at: https://is.muni.cz/th/ub78x/. Supervisor: Miloš Jakubíček.

Rauh, C. & Schwalbach, J. (2020). The ParlSpeech V2 data set: Full-text corpora of 6.3 million parliamentary speeches in the key legislative chambers of nine representative democracies. Available at: https://doi.org/10.7910/DVN/L4OAKN.

Zorrilla-Agut, P. & Fontenelle, T. (2019). IATE 2: Modernising the EU's IATE terminological database to respond to the challenges of today's translation world and beyond. *Terminology*, 25(2), pp. 146–174.

The DICI-A: A Learner Dictionary of Italian Collocations

Stefania Spina[1], Fabio Zanda[2], Irene Fioravanti[1], Luciana Forti[1],

Damiano Perri[2], Osvaldo Gervasi[2]

[1] Università per Stranieri di Perugia

[2] Università degli Studi di Perugia

E-mail: stefania.spina@unistrapg.it, fabio.zanda@unistrapg.it, irene.fioravanti@unistrapg.it, luciana.forti@unistrapg.it, damiano.perri@unipg.it, osvaldo.gervasi@unipg.it

Abstract

In this presentation we describe the DICI-A (Dizionario delle collocazioni italiane per apprendenti), a new learner dictionary of Italian collocations.

The DICI-A includes ca. 11,000 collocations belonging to six syntactic relations: i. Verb + Direct object (*mantenere una promessa*, 'to keep a promise'); ii. Adjective + Noun/Noun + Adjective, where the adjective is a modifier before or after a noun (*brutta avventura*, 'bad adventure'; *tempo libero*, 'free time'); iii. Verb + Adjective (*stare zitto*, 'to stay quiet'); iv. Verb + Adverb, (*fare presto*, 'to hurry up'); v. Adverb + Adjective (*altamente positivo*, 'highly positive'); and vi. Noun + Noun (*parco divertimenti*, 'amusement park').

In the context of Italian phraseological lexicography, in which three different monolingual collocation dictionaries have been published in the last 15 years (Urzì 2009; Tiberii 2012; Lo Cascio 2013), the DICI-A is a lexicographic resource that brings an important added value, since none of the existing dictionaries were specifically aimed at L2 learners, and none were created according to strictly corpus-based criteria.

The presentation will describe the following features of the DICI-A, resulting from methodological choices made during its development:

- it is a corpus-based dictionary: collocations were extracted from an Italian written and spoken reference corpus (Author et al. under review), by integrating measures of frequency and dispersion with association measures (Gablasova et al. 2017; Gries 2024) of exclusivity (Mutual Information; Evert 2005) and strength of association (LogDice; Rychlý 2008);

- the automatically extracted collocations were filtered through a two-step process: a validation against two of the three existing collocation dictionaries, and a human assessment performed by six linguists specialised in phraseology;

- as a dictionary targeted at learners, each entry of the final 11,000 collocational list was assigned to a specific proficiency level (A: base; B; intermediate; and C: advanced)

according to the *Common European Framework of Reference* (Council of Europe 2020), by combining different criteria, such as the rank of collocations in a frequency list, their internal composition, their use by learners at different proficiency levels, attested in a learner corpus of Italian (Author et al. 2023), and their domain of use (La Russa et al. 2023);

- definitions and examples for each of the collocational entries were obtained using Generative AI (Ptasznik et al. 2024): a specific prompt provided through the ChatGPT 4o API interface was found to be effective in producing definitions and examples easily understandable by learners, even at low proficiency levels, as demonstrated by two ad hoc tests (Author et al. 2025).

The DICI-A will be publicly available from the end of 2025 in digital format, and searchable through a dedicated web and mobile interface.

Keywords: learner dictionary; collocations; L2 Italian; association measures; CEFR levels; Generative AI

References

Author et al. (2023).

Author et al. (2025).

Author et al. under review

Council of Europe (2020). *Common European Framework of Reference for Languages: Learning, teaching, assessment – Companion volume*, Council of Europe Publishing, Strasbourg

Evert, S. (2005). *The statistics of word cooccurrences: Word pairs and collocations.* Stuttgart, Germany: University of Stuttgart PhD dissertation.

Gablasova, D., Brezina, V. & McEnery, T. (2017). Collocations in corpus-based language learning research: Identifying, comparing, and interpreting the evidence. *Language Learning* 67(S1), pp. 155–179.

Gries, S. Th. (2024). *Frequency, dispersion, association, and keyness: Revising and tupleizing corpus-linguistic measures.* Studies in Corpus Linguistics, vol. 115. Amsterdam: John Benjamins.

La Russa, F., D'Alesio, V. & Suadoni, A. (2023). Designing a Corpus-Based Syllabus of Italian Collocations: Criteria, Methods and Procedures. *Revue Roumaine de Linguistique, LXVIII*(4), pp. 377–389.

Lo Cascio, V. (2013). *Dizionario Combinatorio Italiano.* Amsterdam: John Benjamins.

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.

Rychlý, P. (2008). A lexicographer-friendly association score. In Petr Sojka & Aleš

Horák (eds.), *RASLAN 2008: Recent Advances in Slavonic Natural Language Processing. Proceedings of the Second Workshop on RASLAN, Karlova Studánka, Czech Republic, December 5–7, 2008*, pp. 6–9. Brno: Masaryk University.

Tiberii, P. (2012). *Dizionario delle collocazioni: Le combinazioni delle parole in italiano.* Bologna: Zanichelli.

Urzě, F. (2009). *Dizionario delle Combinazioni Lessicali.* Lussemburgo: Convivium.

Automatic Detection of Word Sense Shift from Corpus Data

Ondřej Herman

Lexical Computing, Brno, Czech Republic
Natural Language Processing Centre, Masaryk University, Brno, Czech Republic
E-mail: ondrej.herman@sketchengine.eu

Abstract

Language evolves continuously, rendering static dictionaries quickly outdated. While previous research has addressed the automatic detection of new words, identifying subtler semantic changes in existing words remains a challenge. In this work, we propose a robust, language-independent methodology for the automatic detection of word sense shifts using diachronic corpus data. Our approach builds on the Adaptive Skip-Gram algorithm for word sense induction, enabling us to model polysemy directly from raw text without reliance on external sense inventories.

We calculate the temporal distribution of induced senses and apply trend estimation techniques—specifically linear regression and the Theil–Sen estimator—to detect statistically significant shifts. This two-stage architecture decouples sense induction from trend analysis, increasing overall robustness and interpretability. Unlike traditional methods in lexical semantic change detection, which often target dramatic historical shifts, our method is designed to detect emerging or evolving senses over shorter timescales using large web corpora.

We evaluate our method on Timestamped corpora in English and Czech and present several examples of detected sense shifts. The results demonstrate the feasibility of scalable, automatic sense shift detection and its potential applications in lexicography and linguistic research.

Keywords: word sense induction; neologisms; trends

References

Bartunov, S., Kondrashkin, D., Osokin, A. & Vetrov, D. (2016). Breaking sticks and ambiguities with adaptive skip-gram. In *Artificial Intelligence and Statistics*, pp. 130–138.

Cook, P., Lau, J.H., Rundell, M., McCarthy, D. & Baldwin, T. (2013). A lexicographic appraisal of an automatic approach for detecting new word senses. *Proceedings of eLex*, pp. 49–65.

Fiser, D. & Ljubesic, N. (2018). Distributional modelling for semantic shift detection. *International Journal of Lexicography*, 32(2), pp. 163–183.

Frermann, L. & Lapata, M. (2016). A bayesian model of diachronic meaning change. *Transactions of the Association for Computational Linguistics*, 4, pp. 31–45.

Gulordava, K. & Baroni, M. (2011). A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus. In *Proceedings of the GEMS 2011 workshop on geometrical models of natural language semantics*, pp. 67–71.

Herman, O. & Jakubíček, M. (2024). ShadowSense: a Multi-annotated Dataset for Evaluating Word Sense Induction. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 14763–14769.

Herman, O., Jakubíček, M., Kraus, J. & Suchomel, V. (2025). From Word of the Year to Word of the Week: Daily-updated Monitor Corpora for 25 Languages. *Electronic lexicography in the 21st century. Proceedings of the eLex 2025 conference*.

Herman, O. & Kovar, V. (2013). Methods for Detection of Word Usage over Time. In *Seventh Workshop on Recent Advances in Slavonic Natural Language Processing, RASLAN 2013*. Brno: Tribun EU, pp. 79–85.

Kahmann, C., Niekler, A. & Heyer, G. (2017). Detecting and assessing contextual change in diachronic text documents using context volatility. Available at: https://doi.org/10.48550/arXiv.1711.05538.

Kilgarriff, A., Baisa, V., Busta, J., Jakubicek, M., Kovar, V., Michelfeit, J., Rychly, P. & Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, 1(1), pp. 7–36.

Kulkarni, V., Al-Rfou, R., Perozzi, B. & Skiena, S. (2015). Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, pp. 625–635.

Michelfeit, J., Pomikálek, J. & Suchomel, V. (2014). Text Tokenisation Using unitok. In A. Horák & P. Rychlý (eds.) *RASLAN 2014*. Brno, Czech Republic: Tribun EU, pp. 71–75.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S. & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems,* pp. 3111–3119.

Mitra, S., Mitra, R., Maity, S.K., Riedl, M., Biemann, C., Goyal, P. & Mukherjee, A. (2015). An automatic approach to identify word sense changes in text media across timescales. *Natural Language Engineering*, 21(5), pp. 773–798.

Nimb, S., Sřrensen, N.H. & Lorentzen, H. (2020). Updating the dictionary: Semantic change identification based on change in bigrams over time. *Slovenščina 2.0: empirične, aplikativne in interdisciplinarne raziskave*, 8(2), pp. 112–138.

Pomikálek, J. (2011). *Removing boilerplate and duplicate content from web corpora*. Ph.D. thesis, Masaryk university, Faculty of informatics, Brno, Czech Republic.

Rychly, P. (2007). Manatee/Bonito–A Modular Corpus Manager. *RASLAN 2007 Recent Advances in Slavonic Natural Language Processing*, p. 65.

Sagi, E., Kaufmann, S. & Clark, B. (2009). Semantic density analysis: Comparing word meaning across time and phonetic space. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*. Association for Computational Linguistics, pp. 104–111.

Schlechtweg, D. (2023). *Human and computational measurement of lexical semantic change*. Ph.D. thesis, Universität Stuttgart.

Schlechtweg, D., McGillivray, B., Hengchen, S., Dubossarsky, H. & Tahmasebi, N. (2020). SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection. In A. Herbelot, X. Zhu, A. Palmer, N. Schneider, J. May & E. Shutova (eds.) *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. Barcelona (online): International Committee for Computational Linguistics, pp. 1–23.

Schmid, H. (2013). Probabilistic part-of-speech tagging using decision trees. In *New methods in language processing*. Routledge, pp. 154–164.

Tahmasebi, N. & Dubossarsky, H. (2023). Computational modeling of semantic change. Available at: https://arxiv.org/abs/2304.06337.

Tahmasebi, N. & Risse, T. (2017). On the uses of word sense change for research in the digital humanities. In *International Conference on Theory and Practice of Digital Libraries*. Springer, pp. 246–257.

Trampus, M. & Novak, B. (2013). Internals of an aggregated web news feed. In *15th Multiconference on Information Society,* pp. 221–224.

Wittgenstein, L. (1953). *Philosophical Investigations*. New York, NY, USA: Wiley-Blackwell.

Yao, Z., Sun, Y., Ding, W., Rao, N. & Xiong, H. (2018). Dynamic word embeddings for evolving semantic discovery. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. ACM, pp. 673–681.

Zamora-Reina, F.D., Bravo-Marquez, F. & Schlechtweg, D. (2022). LSCDiscovery: A shared task on semantic change discovery and detection in Spanish. In N. Tahmasebi, S. Montariol, A. Kutuzov, S. Hengchen, H. Dubossarsky & L. Borin (eds.) *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*. Dublin, Ireland: Association for Computational Linguistics, pp. 149–164. Available at: https://aclanthology.org/2022.lchange-1.16/.

Do dictionary users prefer definitions by lexicographers or by LLM-s?

Maria Tuulik[1], Ene Vainik[1], Margit Langemets[1], Eleri Aedmaa[1], Lydia Risberg[1,2], Esta Prangel[1], Kristina Koppel[1], Sirli Zupping[1]

[1] Institute of the Estonian Language
[2] University of Tartu
E-mail: maria.tuulik@eki.ee, Ene.Vainik@eki.ee, margit.langemets@eki.ee, eleri.aedmaa@eki.ee, lydia.risberg@gmail.com, esta.prangel@eki.ee, kristina.koppel@eki.ee, sirli.zupping@eki.ee

Abstract

The use of corpora is well established in lexicography, also in Estonia, but since the analysis of corpus data and the post-editing of automatically generated data from the corpus is labour-intensive, the use of large language models (LLMs) has led to growing interest in lexicography (e.g., Evert et al. 2024; Kosem, Gantar et al. 2024; Tiberius et al. 2024). In 2024, the Institute of the Estonian Language launched a project in which we explore how LLMs can assist in compiling dictionary entries (e.g., definitions, register labels, examples).

In the first year, we tested whether LLMs can help lexicographers in the task of explaining word meanings in Estonian, a language with around 1 million speakers and underrepresented in LLMs. The results showed that lexicographers rated 85% of the GPT-4o (highest rated LLM in the study) generated meaning descriptions as useful or somewhat useful for their work. While our first study focused on lexicographers' preferences and requirements for LLM-generated definitions, in the current study we concentrate on users' preferences and requirements for both, LLM-generated and lexicographer-compiled definitions.

According to a survey conducted in 2023 (Langemets et al. 2024: 750-751), the Estonian Language Institute's language portal *Sõnaveeb* (Koppel et al. 2019) is searched most for information on meanings. This coincides with the results of a pan-European study (Kosem et al., 2019), according to which meanings in general are the most searched units in dictionaries. However, both studies were carried out before the wider use of LLMs. No research has been carried out on the Estonian language to investigate whether and how preferences for obtaining information about meanings have changed with the increasing use of LLMs. In the presentation, we will introduce the results of a survey carried out among the users of *Sõnaveeb*, where LLM generated definitions were presented side-by-side to lexicographer compiled definitions, and users had to mark their preference and list the reasons for it. The evaluation is conducted blindly, with

users not being informed which explanation is human-made. The lexicographic meaning descriptions used in the survey are the definitions from the the *EKI Combined Dictionary* (Tavast et al. 2020), which is the backbone of *Sõnaveeb* and presents a monolingual detailed description of meaning that defines the content of the concept as exhaustively as possible. Words from different parts of speech and with varying degrees of polysemy were included in the study.

We tested the following LLMs: GPT-4o, o1mini, Claude 3 Opus, Claude 3.5 Sonnet, Gemini 1.5 Pro, Gemini 2.0 ja Euro LLM. Based on expert evaluations, the best-performing model was selected for the final user test. In the presentation, we introduce the tested prompts and examine how users' dictionary and LLM usage habits relate to their preferences. But mainly, how do users rate the LLM-generated definitions, and do they prefer them to the ones lexicographers compiled? What do lexicographers still do better than LLMs, and what, intriguingly, do users believe LLMs do better than lexicographers?

Keywords: definitions; LLMs; dictionary users; Estonian language

References

Evert, S., Ganslmayer, C. & Rink, C. (2024). Multi-Level Analysis as a Systematic Approach to Evaluating the Quality of AI-Generated Dictionary Entries. *Proceedings of the XXI EURALEX International Congress*, pp. 317–334.

Koppel, K., Tavast, A., Langemets, M. & Kallas, J. (2019). Aggregating dictionaries into the language portal Sõnaveeb: issues with and without a solution. In: Kosem, I., Zingano Kuhn, T., Correia, M., Ferreria, J. P., Jansen, M., Pereira, I., Kallas, J., Jakubíček, M., Krek, S. & Tiberius, C. (Ed.). Proceedings of the eLex 2019 conference. 1-3 October 2019, Sintra, Portugal, pp. 434−452. Brno: Lexical Computing CZ, s.r.o.

Kosem, I., Lew, R., Müller-Spitzer, C., Ribeiro Silveira, M., Wolfer, S. et al. (2019). The Image of the Monolingual Dictionary Across Europe. Results of the European Survey of Dictionary use and Culture. *International Journal of Lexicography*, 32(1), pp. 92–114.

Kosem, I., Gantar, P., Arhar Holdt, Š., Gapsa, M., Zgaga, K. & Krek, S. (2024). AI in Lexicography at the University of Ljubljana: Case Studies. *XXI EURALEX International Congress. Lexicography and Semantics*, Cavtat, Croatia.

Langemets, M., Risberg, L. & Algvere, K. (2024). To Dream or Not to Dream About 'Correct' Meanings? Insights into the User Experience Survey. Lexicography and Semantics. *Proceedings of the XXI EURALEX International Congress*. Ed. Kristina Š. Despot, Ana Ostroški Anić, Ivana Brač. Institute for the Croatian Language, pp. 741–760.

Tavast, A., Koppel, K., Langemets, M. & Kallas, J. (2020). Towards the superdictionary: layers, tools and unidirectional meaning relations. *Proceedings of*

XIX EURALEX Congress: Lexicography for Inclusion, Vol. I. Ed. Gavriilidou, Z, Mitsiaki, M, Fliatouras, A. Alexandroupolis, Greece: Democritus University of Thrace, pp. 215−223.

Tiberius, C., Heylen, K., Vanroy, B., Vandeghinste, V., de Does, J. & van Doeselaar, J. (2024). LLMs and Evidence-Based Lexicography: Pilot Studies at INT. *XXI EURALEX International Congress. Lexicography and Semantics*, Cavtat, Croatia. Available at: https://euralex.jezik.hr/wp-content/uploads/2021/09/EURALEX_2024_Full_Programme_chairs_FINAL.pdf.

Retention of English words from interaction with dictionaries and GenAI Chatbots

Robert Lew[1], Bartosz Ptasznik[2]

[1] Adam Mickiewicz University

[2] University of Warmia and Mazury in Olsztyn, Poland

E-mail: rlew@amu.edu.pl, bartosz.ptasznik@uwm.edu.pl

Abstract

The public release of ChatGPT in late 2022 made an impact on many professional domains. Notwithstanding the many controversies surrounding Generative Artificial Intelligence (GenAI), such as ethics, copyright, accountability, or ecology, we need to acknowledge an important and relevant feature of Large Language Models and chatbot systems built around them: their ability to produce mostly natural-sounding, smooth English prose. This ability makes AI Chatbots an attractive option in the learning (and teaching) of English, and thus a serious competitor to dictionaries seen as traditional learning (and teaching) aids, especially when it comes to vocabulary: the natural focus of lexicography and dictionaries. Effective use of dictionaries requires specific dictionary skills (e.g. Nesi, 1999), whereas AI Chatbots are generally believed to be straightforward and quick to use. A few recent studies have indeed found that ChatGPT may result in better student performance on English vocabulary tasks compared to traditional bilingual and monolingual dictionaries, at least for production tasks, if not always in reception (Lew et al., 2024; Ptasznik et al., 2024; Rees and Lew, 2024). These studies focused on immediate success, but we are not aware of any studies that would investigate vocabulary retention. It is quite possible that the ease and speed with which Chatbots facilitate the immediate completion of language-related tasks might not promote learning (a concern in fact we often hear from AI critics).

In our eLex 2025 presentation, we report on two ongoing studies looking beyond immediate success and at delayed retention. Both studies tested the reception and production of infrequent and semantically opaque English phrasal verbs (20 in Study One, 19 in Study Two). Polish students majoring in English were randomized to one of three tools and completed reception and production tasks focuses on phrasal verbs. Two to three weeks later they were re-tested, but now without access to any lexical tools. Study One tested the bilingual dictionary bab.la, the monolingual Collins Online Dictionary, and ChatGPT and found modest but significant and similar learning gains with all three tools in a reception task. For delayed production, ChatGPT was the only tool to result in significant learning. Study 2 used a larger sample (223 participants) and two different chatbots as well as the bilingual dictionary diki.pl which had been found effective in an earlier study (Lew et al., 2024). In delayed reception tests, the

bilingual dictionary significantly outperformed both MS Copilot and Gemini, whereas for production, no significant differences were found between any of the tools, just an effect of the year of study. Our general tentative conclusion is that completing lexically oriented tasks with the help of AI chatbots does not seriously disadvantage longer-term vocabulary retention, compared to dictionaries.

Keywords: English; vocabulary learning; vocabulary acquisition; AI Chatbots

References

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. *Humanit Soc Sci Commun*, 11, 1324. Available at: https://doi.org/10.1057/s41599-024-03775-y.

Nesi, H. (1999). The specification of dictionary reference skills in higher education, in: Hartmann, R.R.K. (Ed.) *Dictionaries in Language Learning. Recommendations, National Reports and Thematic Reports from the Thematic Network Project in the Area of Languages, Sub-Project 9: Dictionaries.* Freie Universität Berlin, Berlin, pp. 53–67.

Ptasznik, B., Wolfer, S. & Lew, R. (2024). A Learners' Dictionary Versus ChatGPT in Receptive and Productive Lexical Tasks. *International Journal of Lexicography,* 37, pp. 322–336. Available at: https://doi.org/10.1093/ijl/ecae011.

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, pp. 50–74. Available at: https://doi.org/10.1093/ijl/ecad030.

Compiling bilingual dictionaries: AI-Assisted translation of Italian Multiword Expressions into English and French

Annalisa Greco[1], Matteo Delsanto[2], Andrea Di Fabio[2], Lorenzo Mori[2], Cristina Onesti[1], Daniele Paolo Radicioni[2], Calogero Jerik Scozzaro[2]

[1] Università degli Studi di Torino, Dipartimento di Lingue e Letterature straniere e Culture Moderne, via Sant'Ottavio, 18, 10124, Torino

[2] Università degli Studi di Torino, Dipartimento di Informatica, Corso Svizzera, 185, 10149, Torino

E-mail: annalisa.greco@unito.it, matteo.delsanto@unito.it, andrea.difabio@unito.it, lorenzo.mori31@edu.unito.it, cristina.onesti@unito.it, daniele.radicioni@unito.it, calogerojerik.scozzaro@unito.it

Abstract

The present research explores the use of large language models (LLMs) in digital lexicography, specifically for translating Italian multiword expressions (MWEs) into English and French.

The study aims to assess the capability of contemporary LLMs in providing accurate and reliable translation equivalents, examples and definitions of Italian MWEs into English and French, while also evaluating the need for expert validation in refining AI-generated lexicographic resources. We seek to develop a digital resource tailored for language learners, offering frequently attested translations.

Methodologically, 120 expressions were evaluated by human experts and compared across two LLMs (Gemini 2.0 Flash and Mistral-Large-2411) using different metrics aimed at assessing including correctness, accuracy and contextual suitability, along with the capacity to produce meaning explanations and usage examples. Results show that English translations received higher expert ratings than French ones, with high correlation between human and AI evaluations in the case of English, and significantly lower agreement in the case of French translations. The findings indicate that LLMs provide generally reliable translations, though expert oversight remains crucial.

Keywords: multiword expressions; large language models; AI-assisted translation; bilingual dictionaries; dictionary writing system/dictionary-making process

References

Books:

Burger, H. (2010). Phraseologie. Eine Einführung am Beispiel des Deutschen,4., neu bearbeitete Auflage (Grundlagen der Germanistik 36). Berlin: ErichSchmidt.

Giouli, V. & Barbu Mititelu, V. (2024) (eds.). *Multiword expressions in lexical resources: Linguistic, lexicographic, and computational perspectives. (Phraseology and Multiword Expressions 6).* Berlin: Language Science Press. Available at: https://langsci-press.org/catalog/book/440.

Hjelmslev, L. (1968). *I fondamenti della teoria del linguaggio*. Torino: Einaudi.

Konecny, C. (2010). Kollokationen. Versuch einer semantisch-begrifflichen Annäherung und Klassifizierung anhand italienischer Beispiele. München: Martin Meidenbauer [Forum Sprachwissenschaften; 8].

Marello, C. (1996). *Le parole dell'italiano. Lessico e dizionari*. Bologna: Zanichelli.

Ramisch, C. (2023). *Multiword expressions in computational linguistics*. Computer Science [cs]. Aix Marseille Université (AMU).

Book sections:

Abel, A. (2012). Dictionary writing systems and beyond. In S. Granger & M. Paquot (eds.), *Electronic Lexicography*. Oxford, online edn, Oxford Academic, 24 Jan. 2013. Available at: doi.org/10.1093/acprof:oso/9780199654864.003.0005.

Calzolari, N. *et al.* (2002). Towards best practice for multiword expressions in computational lexicons. In M. G. Rodríguez & C. P. S. Araujo (eds.), *Towards Best Practice for Multiword Expressions in Computational Lexicons*. LREC, pp. 1934–1940.

De Mauro, T. (1999). *Introduzione*. In GRADIT 1999-2007, vol. 1º, pp. VII-XLII.

Faloppa, F. (2011). Modi di dire. In S. Raffaele (ed.), *Enciclopedia dell'Italiano (EncIt)*, Roma, Istituto della Enciclopedia italiana. Available at: https://www.treccani.it/enciclopedia/modi-di-dire_(Enciclopedia-dell'Italiano)/.

Masini, F. (2019). Multi-Word Expressions and Morphology. In *Oxford Research Encyclopaedia of Linguistics*. Oxford, Oxford University Press.

Voghera, M. (2004). Polirematiche. In M. Grossmann & F. Rainer (eds.), *La formazione delle parole in italiano*. Tübingen: Niemeyer, pp. 56–69.

Paper in conference proceedings:

Garcia, M. *et al.* (2019). Towards the Automatic Construction of a Multilingual Dictionary of Collocations using Distributional Semantics. In I. Kosem & T. Zingano Kuhn (eds.). *Electronic lexicography in the 21st century. Proceedings of the eLex 2019 Conference*, 1-3 October 2019, Sintra, Portugal. Brno: Lexical Computing CZ, s.r.o., pp. 747–762. Available at: https://elex.link/elex2019/wp-

content/uploads/2019/09/eLex_2019_42.pdf.

Lee, N. et al. (2025). Evaluating the Consistency of LLM Evaluators. In *Proceedings of the 31st International Conference on Computational Linguistics,* pp. 10650–10659.

Orenha-Ottaiano, A. (2017). The Compilation of an Online Corpus-Based Bilingual Collocations Dictionary: Motivations, Obstacles and Achievements. In *Proceedings of E-Lex Conference 2017*, Leiden, The Netherlands, pp. 458–473. Available at: https://elex.link/elex2017/wpcontent/uploads/2017/09/paper27.pdf.

Orenha-Ottaiano, A. *et al.* (2021). Corpus-based Methodology for an Online Multilingual Collocations Dictionary: First Steps. In Kosem, I., Cukr, M., Jakubíček, M., Kallas, J., Krek, S. & Tiberius, C. (eds.), *Proceedings of Electronic Lexicography in the 21st Century Conference*, 2021-July, pp. 1–28. Available at: https://elex.link/elex2021/wp-content/uploads/eLex_2021-proceedings_compressed.pdf.

Sag, I. *et al.* (2002). Multiword Expressions: a Pain in the Neck for NLP. In Gelbukh, A. (ed.), *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, Lecture Notes in Computer Science, vol. 2276. Springer, Berlin, Heidelberg. Available at: https://doi.org/10.1007/3-540-45715-1_1, pp. 1–15.

Journal articles:

Gantar, P. *et al.* (2019). Multiword expressions: between lexicography and NLP. In *International Journal of Lexicography*, Vol. 32, n. 2, pp.138–162.

Greco, A. (in preparation), "Una rivisitazione di alcune categorie di 'combinazioni di parole': alcuni criteri lessicografici per la compilazione del Dizionario Nativo Digitale (DND)".

Lew, R. (2024). Dictionaries and lexicography in the AI era. *Humanit Soc Sci Commun* 11, 426. Available at: https://doi.org/10.1057/s41599-024-02889-7.

Lew, R. *et al.* (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), pp. 1–10.

Mohammed, T. A. (2025). Evaluating Translation Quality: A Qualitative and Quantitative Assessment of Machine and LLM-Driven Arabic–English Translations. Information, 16(6), 440.

Uchida, S. (2024). Using early LLMs for corpus linguistics: Examining ChatGPT's potential and limitations. *Applied Corpus Linguistics*, 4(1), 100089.

Villavicencio, A. *et al.* (2005) (eds). Special issue on multiword expressions: Having a crack at a hard nut. In *Computer Speech & Language*, Volume 19, Issue 4, pp. 365–377.

Websites:

The Collins Online Dictionary. Accessed at: https://www.collinsdictionary.com/.

Larousse. Accessed at: https://www.larousse.fr/.

Le Grand Robert. Accessed at: https://www.lerobert.com/.

Le Petit Robert. Accessed at: https://www.lerobert.com/.
Le Robert Dico en ligne. Accessed at: https://dictionnaire.lerobert.com/.
The Merriam-Webster. Accessed at: https://www.merriam-webster.com/.
The *Oxford Learner's Dictionaries* (learner's dictionaries). Accessed at: http://www.oxfordlearnersdictionaries.com.
Le Trésor de la langue française (TLFi). Accessed at: http://atilf.atilf.fr/.

Dictionaries:

De Mauro Internazionale = De Mauro, T., *Dizionario di italiano* (dizionario.internazionale.it/).
Lapucci, C. (2007). *Dizionario dei proverbi italiani*. Milano: Mondadori.
Russo, D. (2010). *Modi di dire. Lessico italiano delle collocazioni*. Roma: Aracne.
Tiberii, P. (2012). *Dizionario delle collocazioni*. Bologna: Zanichelli.

Ensemble Approach to Lemmatization of Out-of-Vocabulary Lexical Items in Slovak Corpora

Vladimír Benko

Slovak Academy of Sciences, Ľ. Štúr Institute of Linguistics
Panská 26, 811 01 Bratislava, Slovakia
E-mail: vladimir.benko@juls.savba.sk

Abstract

Our paper addresses the problem of lemmatizing out-of-vocabulary lexical items (OOVs) in large Slovak corpora. Using the ensemble approach, the results of statistical guessing provided by some of the available taggers can be attested and/or disambiguated. Considering the large scale of the data and the available hardware, only tools not requiring graphics cards were considered.

Keywords: corpus tagging; lemmatization; OOVs; guessing; ensemble approach

References

Afanasev, I., Glazkova, A., Lyashevskaya, O., Morozov, D., Smal I., and Vlasova, N. (2025). Rubic2: Ensemble Model for Russian Lemmatization. *Slavic NLP Workshop*. Vienna, 31 July 2025.

Benko, V. (2014). Aranea: Yet Another Family of (Comparable) Web Corpora. In *Petr Sojka, Aleš Horák, Ivan Kopeček and Karel Pala (Eds.): Text, Speech and Dialogue. 17th International Conference, TSD 2014*, Brno, Czech Republic, September 8-12, 2014. Proceedings. LNCS 8655. Springer International Publishing Switzerland (2014).

Benko, V. (2024). The Aranea Corpora Family: Ten+ Years of Processing Web-Crawled Data. In *Lecture Notes in Computer Science: Text, Speech, and Dialogue.* Proceedings, Part 1. - Heidelberg : Springer, 2024, vol. 15048, pp. 55–70.

Dell'Orletta, F. (2009). Ensemble system for Part-of-Speech tagging. *Evalita 2009 POS Closed Task.* (unpublished?)

Garabík, R. – Mitana, D. (2023). Analysing Accuracy of Slovak Language Lemmatization and MSD Tagging. In: *Slovenská reč*, 88/2, 129–140.

Hajič, J. (2004) Disambiguation of Rich Inflection: Computational Morphology of Czech. Karolinum Press, 2004.

Jongejan, B, and Dalianis, H. (2009). Automatic training of lemmatization rules that handle morphological changes in pre-, in- and suffixes alike. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International*

Joint Conference on Natural Language Processing of the AFNLP. Suntec, Singapore: Association for Computational Linguistics, pp. 145–153.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing,* Manchester.

Sourada, T., Straková, J., Rosa, R. (2024). OOVs in the Spotlight: How to Inflect them? *LREC-COLING 2024*, pp. 12455–12466, 20-25 May, 2024.

Spoustová, D. „johanka“, Hajič, J., Raab, J., Spousta., M. (2009). Semi-Supervised Training for the Averaged Perceptron POS Tagger. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL)*, pp. 763–771, Athens, Greece, March 2019. Association for Computational Linguistics.

Straka, M., Hajič J., Straková J. (2016). UDPipe. Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC), Portorož, Slovenia.*

Documenting the Final Days of Monolingual English Learners' Dictionaries Using the Archived Web

Geraint Paul Rees

Department of Translation and Language Sciences, Universitat Pompeu Fabra, Barcelona
E-mail: geraintpaul.rees@upf.edu

Abstract

Online dictionaries have many advantages over their physical counterparts. However, the ephemeral nature of web content means that they are often changed without notice and no ostensible record of what came before remains. This makes research on historical online dictionaries difficult and perhaps explains why, while the history of printed monolingual English learners' dictionaries (MELDs) has been comprehensively explored, studies of online dictionaries have tended to take a cross-sectional rather than longitudinal view. This is not ideal since it means that a large period of MELD history is yet to be explored. Moreover, given recent predictions of the decline of MELDs, as we know them, in light of developments with AI chatbots and other digital tools, this gap is all the more significant. In an attempt to remedy this situation, this study applies Brügger's (2018) framework for archived web research to explore the feasibility of using the web archive, *the Wayback Machine,* to trace the development of websites that give, or have given, access to 'the big five' MELDs. Some key challenges of using archived web material to conduct lexicographic research are discussed along with suggestions for potential solutions.

Keywords: digital dictionaries; monolingual English learner's dictionaries (MELDs); web archives; Internet Archive; history of lexicography

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, pp. 545–560. Available at: https://doi.org/10.1007/s10209-023-01031-9.

Brügger, N. (2015). A brief history of Facebook as a media text: The development of an empty structure. *First Monday*, 20(5). Available at: https://firstmonday.org/ojs/index.php/fm/article/view/5423.

Brügger, N. (2018). *The Archived Web: Doing History in the Digital Age*. The MIT Press. Available at: https://doi.org/10.7551/mitpress/10726.001.0001.

Cambridge Dictionaries. (2000, December 2). *All-time top 20*. Available at: https://web.archive.org/web/20001202072500/http://dictionary.cambridge.org/t

op20.htm.

Cowie, A.P. (1999). *English Dictionaries for Foreign Learners: A History*. Clarendon Press. Available at: https://global.oup.com/academic/product/english-dictionaries-for-foreign-learners-9780198235064.

De Schryver, G. (2003). Lexicographers' dreams in the electronic-dictionary age. *International Journal of Lexicography*, 16(2), pp. 143–199. Available at: https://doi.org/10.1093/ijl/16.2.143.

De Schryver, G.-M. (2023). The future of the dictionary. In E. Finegan & M. Adams (eds.) *The Cambridge Handbook of the Dictionary*. Cambridge University Press.

Dziemianko, A. (2015). Colours in online dictionaries: A case of functional labels. *International Journal of Lexicography*, 28(1), pp. 27–61. Available at: https://doi.org/10.1093/ijl/ecu028.

Dziemianko, A. (2018). Electronic dictionaries. In P. Fuertes-Olivera (ed.) *The Routledge Handbook of Lexicography*, pp. 663–683. Routledge.

Dziemianko, A. (2019). The role of online dictionary advertisements in language reception, production, and retention. *ReCALL*, 31(1), pp. 5–22. Available at: https://doi.org/10.1017/S0958344018000149.

Dziemianko, A. (2020). Smart advertising and online dictionary usefulness. *International Journal of Lexicography*, 33(4), pp. 377–403. Available at: https://doi.org/10.1093/ijl/ecaa017.

Frankenberg-Garcia, A., Lew, R., Rees, G.P., Roberts, J., Sharma, N. & Butcher, P. (2019, September). *Collocations in e-Lexicography: lessons from Human computer interaction research* [Workshop presentation]. Pre-conference workshop on collocations at eLex 2019, Sintra.

Fuertes-Olivera, P.A. & Tarp, S. (2020). A window to the future: Proposal for a lexicography-assisted writing assistant. *Lexicographica*, 36(2020), pp. 257–286. Available at: https://doi.org/10.1515/lex-2020-0014.

Hao, J., Xu, H. & Hu, H. (2022). A multimodal communicative approach to the analysis of typography in online English learner's dictionaries. *International Journal of Lexicography*, 35(2), pp. 234–260. Available at: https://doi.org/10.1093/ijl/ecab031.

Hirtle, P.B. (2003, November). *Digital Preservation and Copyright*. Stanford University. Available at: https://fairuse.stanford.edu/2003/11/10/digital_preservation_and_ copyr/.

Internet Archive. (2025). *Wayback Machine*. Available at: https://archive.org/web/.

Jessen, I.B. (2010). The aesthetics of web advertising: Methodological mplications for the study of genre development. In N. Brügger (ed.) *Web History*, pp. 257–277. Peter Lang.

Lew, R. (2024). Dictionaries and lexicography in the AI era. *Humanities and Social Sciences Communications*, 11(1), pp. 1–8. Available at: https://doi.org/10.1057/s41599-024-02889-7.

Lew, R. & Wolfer, S. (2024). CEFR vocabulary level as a predictor of user interest in English Wiktionary entries. *Humanities and Social Sciences Communications*,

11(1), pp. 340. Available at: https://doi.org/10.1057/s41599-024-02838-4.

Nesi, H. (2024). Are we witnessing the death of dictionaries? *Ibérica*, *47*, Article 47. Available at: https://doi.org/10.17398/2340-2784.47.7.

Preston-Kendal, D. (2025). Dictionaries in the web of Alexandria: On the dangerous fragility of digital publication. In G. Williams, M.L. Meur, & A.E. Peláez (eds) *West Meets East: Papers in Historical Lexicography and Lexicology from Across the Globe*, pp. 31–41. Language Science Press. Available at: https://doi.org/10.5281/zenodo.15394473.

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: Preparation, Exploitation, Analysis,* pp. 32–54. Routledge.

Rees, G.P. (2022). Using corpora to write dictionaries. In A. O'Keeffe & M. McCarthy (eds.) *The Routledge Handbook of Corpus Linguistics* (Second edition), pp. 387–404. Routledge.

Rees, G.P. (2023). Online dictionaries and accessibility for people with visual Impairments. *International Journal of Lexicography*, 36(2), pp. 107–132. Available at: https://doi.org/10.1093/ijl/ecac021.

Rees, G.P. (2024). Academic word families in online English dictionaries. *Lexikos*, 34, pp. 437–468. Available at: https://doi.org/10.5788/34-1-1947.

Rees, G.P. (2025a). Exploring which aspects of an online monolingual learners' dictionary can be investigated using the archived web. *International Journal of Lexicography*, ecaf020. Available at: https://doi.org/10.1093/ijl/ecaf020.

Rees, G.P. (2025b). Multimedia in dictionaries. In H. Nesi & P. Milin (eds) *International Encyclopedia of Language and Linguistics* (3rd edn). Elsevier. Available at: https://doi.org/10.1016/B978-0-323-95504-1.00685-2.

Rees, G.P. (2025c). Making dictionary content accessible for people with visual Impairments. *Lexikos*, 35(2), pp. 165–184. Available at: https://doi.org/10.5788/35-2-2077.

Rees, G.P. & Frankenberg-Garcia, A. (2025a). Writing aids for DDL. In L. McCallum & D. Tafazoli (eds.) *The Palgrave Encyclopedia of Computer-Assisted Language Learning* (pp. 1–6). Springer Nature Switzerland. Available at: https://doi.org/10.1007/978-3-031-51447-0_55-1.

Rees, G.P. & Frankenberg-Garcia, A. (2025b). Dictionaries embedded: writing assistants and other tools. In *Reference Module in Social Sciences*. Elsevier. Available at: https://doi.org/10.1016/B978-0-323-95504-1.00876-0.

Roberts, J.C., Butcher, P.W.S., Lew, R., Rees, G.P., Sharma, N. & Frankenberg-Garcia, A. (2020). Visualising Collocation for Close Writing. In A. Kerren, C. Garth & G.E. Marai (eds.) *EuroVis 2020—Short Papers*, pp. 181–185. The Eurographics Association. Available at: https://doi.org/10.2312/evs.20201069.

Wolfer, S. & Lew, R. (2025). Supplementing CEFR-graded vocabulary lists for language learners by leveraging information on dictionary views, corpus frequency, part-of-speech, and polysemy. *Humanities and Social Sciences Communications*, 12(1),

pp. 1151. Available at: https://doi.org/10.1057/s41599-025-05446-y.

Yamada, S. & Xu, H. (2024). Learner's dictionaries. In E. Finegan & M. Adams (eds) *The Cambridge Handbook of the Dictionary*, pp. 109–130. Cambridge University Press. Available at: https://doi.org/10.1017/9781108864435.007.

Compiling a candidate list of taboo constructions for an under-resourced language

Monique Rabé[1], Martin J. Puttkammer[2], Gerhard B. van Huyssteen[2]

[1] School of Languages, North-West University, Potchefstroom, South Africa
[2] Centre for Text Technology (CTexT), North-West University, Potchefstroom, South Africa
E-mail: Monique.Rabe@nwu.ac.za, Martin.Puttkammer@nwu.ac.za,
Gerhard.VanHuyssteen@nwu.ac.za

Abstract

Taboo-language resources remain scarce for under-resourced languages like Afrikaans – despite their clear relevance for natural language processing (NLP) and applications in artificial intelligence (AI). Although Afrikaans has a long-standing lexicographic tradition, it still lacks an open-access reusable lexical database for the taboo language. One of the most crucial steps in developing a constructional database for taboo language is to identify a candidate list of taboo constructions for potential lexicographic treatment. This paper outlines and tests a range of procedures to compile and refine such a list, with the goal of establishing a replicable methodology for similar work in other under-resourced languages. The methods draw on existing data of different types and corpora representing different registers. However, many entries are either false positives or ambiguous and require validation. Hence, we experiment with various semi-automated modelling techniques. These techniques include refining the candidate list through frequency analyses in corpora, expanding the list through partial corpus matching, and comparing the results against an attested, verified subset of taboo terms.

Keywords: Afrikaans; candidate list; lexical database; taboo language; under-resourced languages

References

Beyer, H.L. & Louw, P.A. (2022). Aspekte van vernuwing in die Afrikaanse leksikografie: Standaard- en pedagogiese woordeboeke as barometer [Aspects of innovation in Afrikaans lexicography: Standard and pedagogical dictionaries as a barometer]. *Lexikos*, 32(3), pp. 25–48. Available at: https://doi.org/https://doi.org/ 10.5788/32-3-1730.

Dekker, L. (1991). "Vloek, skel en vulgariteit: Hantering van sosiolinguisties aanstootlike leksikale items [Swearing, name-calling and vulgarity: Treatment of sociolinguistically offensive lexical items]." *Lexikos*, 1, pp. 52–60. Available at: https://doi.org/https://doi.org/10.5788/1-1-1148.

Harteveld, P. & Van Niekerk, A.E. (1995). Beleid vir die hantering van beledigende en sensitiewe leksikale items in die Woordeboek van die Afrikaanse Taal (WAT) [Policy for the treatment of insulting and sensitive lexical items in the Woordeboek van die Afrikaanse Taal (WAT)]. *Lexikos*, 5(5), pp. 232–248.

Ramos, G., Batista, F., Ribeiro, R., Fialho, P., Moro, S., Fonseca, A., Guerra, R., Carvalho, P., Marques, C. & Silva, C. (2024). A comprehensive review on automatic hate speech detection in the age of the transformer. *Social Network Analysis and Mining*, 14(1), pp. 204. Available at: https://doi.org/10.1007/s13278-024-01361-3.

Van Huyssteen, G.B. (1998). Die leksikografiese hantering van seksuele uitdrukkings in Afrikaans [The lexicographic handling of sexual expressions in Afrikaans]. *South African Journal of Linguistics*, 16(2), pp. 63–71. Available at: https://doi.org/https://doi.org/10.1080/10118063.1998.9724137.

Van Huyssteen, G.B. & Tiberius, C. (2023). Towards a lexical database of Dutch taboo language. 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, eLex 2023*. Brno, Czech Republic. pp. 53–74.

Wiegand, M., Ruppenhofer, J., Schmidt, A. & Greenberg, C. (2018). Inducing a lexicon of abusive words – a feature-based approach. In M. Walker, H. Ji & A. Stent (eds.) *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), NAAC HLT 2018*. New Orleans, Louisiana. pp. 1046–1056.

The Mangalam Dictionary of Buddhist Sanskrit: automating lexicographic data with generative LLMs

Ligeia Lugli

Mangalam Research Center, 2018 Allston Way, Berkeley (CA) USA
E-mail: ligeia.lugli@london.ac.uk

Abstract

This paper reports on recent advancements in the development of the *Mangalam Dictionary of Buddhist Sanskrit*, the first corpus-driven dictionary dedicated to Buddhist Sanskrit. This is a low-resource, historical, and domain-specific language variety instantiated in South Asian Buddhist literature dating from approximately the first millennium CE. The paper focusses on advances in the automation of this dictionary's data with generative Large Language Models (LLMs), with a view to share our solutions with scholars working with other low-resource historical languages. Specific doomed to fail ally, the paper addresses the effectiveness and viability of leveraging latest generation LLMs to automate three tasks that are central to our lexicographic work: semantic annotation of corpus sentences, identification of a headword's semantic prosody in different contexts, and comparison of a headword's synonyms. The paper first evaluates the relative performance of different commercially available models (including GPT 4.1, Sonnet4 and Gemini 2.5) on a semantic tagging task and then details different approaches we experimented with for enriching our corpus with word-sense and semantic prosody tags using LLMs. It concludes with a brief discussion of commercial LLMs' ability to compare Sanskrit synonyms on the basis of corpus sentences.

Keywords: Buddhist Sanskrit; generative LLMs; semantic tagging; historical corpora

References

Edgerton, F. (1953). *Buddhist Hybrid Sanskrit Grammar and Dictionary* (2 vols.). New Haven: Yale University Press.

Guo, X., Ma, F., Dienes, Z., & Graham, S. (2011). "Acquisition of conscious and unconscious knowledge of semantic prosody." *Consciousness and Cognition*, 20(3), pp. 481–492.

Lugli, L. (2021a). Dictionaries as collections of data stories: an alternative post-editing model for historical corpus lexicography. In Iztok Kosem, et al. (eds.). *Post-Editing Lexicography: eLex 2021*, pp. 216–231.

Lugli, L. (2021b). Words or terms? Models of terminology and the translation of

Buddhist Sanskrit vocabulary. In Alice Collett (ed.) *Buddhism and Translation: Historical and Contextual Perspectives*, New York: SUNY, pp. 149–172.

Lugli, L. (2022). Embeddings models for Buddhist Sanskrit. *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC 2022)*, pp. 3861–3871.

Lugli, L., Martinc, M., Pollak, S., Pelicon, A. (2023). Computing the Dharma: NEH White Paper. figshare. Available at: https://doi.org/10.6084/m9.figshare.24065868.v1.

Martinc, M, Pelicon, A, Pollak, S, Lugli, L. (2023). Word Sense Induction on corpus of Buddhist Sanskrit literature. In Medved M. et al (eds), *Proceeding of the eLex 2023 Conference: Silent Lexicography*, pp. 191–205.

McGillivray, B., Kondakova, D., Burman, A., Dell'Oro, F., Bermúdez Sabel, H., Marongiu, P., Márquez Cruz, M. (2022). A new corpus annotation framework for Latin diachronic lexical semantics, *Journal of Latin Linguistics*, vol. 21(1), pp. 47–105.

Monier-Williams, M. (1899). *A Sanskrit-English Dictionary: Etymologically and Philologically Arranged with Special Reference to Cognate Indo-European Languages*. Oxford: The Clarendon Press.

Stewart, Dominic. (2010). Semantic Prosody: A Critical Evaluation. Routledge.

Vatri, A., McGillivray, B. (2018). The Diorisis ancient Greek corpus: Linguistics and literature. *Research Data Journal for the Humanities and Social Sciences*, 3(1), pp. 55–65.

You get it through lexicography: extracting suppressed language from LLMs using lexicographic scenarios as jailbreaking tools

Esra Abdelzaher, Ágoston Tóth

Department of English Linguistics, Institute of English and American Studies
University of Debrecen
E-mail: esra.abdelzaher@gmail.com, toth.agoston@arts.unideb.hu

Abstract

Taboo words present a challenge for a lexicographer to include and describe in a language resource, as they are forms of verbal violence. However, discarding offensive words from general-purpose lexicographic wordlists disregards the representation of an integral part of the mental lexicon. The present study aims at using lexicographic scenarios to jailbreak four GPT variants into the retrieval of offensive words that are frequently used yet undocumented in most lexicographic resources. While Large Language Models (LLMs) can be used to document a headword, the presence of taboo items may prevent these systems from providing an answer. Our results reveal that the type of the model and the lexicographic framing of the extraction task improved the responses of the models and increased the success rate, with the optimal configuration reaching 87.5% success rate. The AI-generated lexicon of offensive words currently contains approximately 250 headwords grouped into gender, age, religion and race categories. The words also vary in their inherently or contextually offensive types. A searchable user-friendly version is accessible through https://arabic-studies.com/Elex/index.html. The main contributions of this lexicon are detecting lexicographically undocumented offensive terms, pointing to the negative context of several headwords and discovering new senses of apparently neutral ones. In addition, LLMs provide very useful morphological, semantic and socio-cultural information in the definitions, despite the inconsistencies and some overgeneralizations in the definitions. Although corpus evidence proved the success of LLMs in detecting offensive words and senses, the automatic evaluation of AI-generated example sentences showed their limited value from a pedagogical perspective.

Keywords: Offensive language; Jailbreak; Prompt engineering; GPT

References

Abdelhakim, M., Liu, B. & Sun, C. (2023). Ar-Pufi: A short-text dataset to identify the offensive messages towards public figures in the Arabian community. *Expert Systems with Applications*, 233, 120888.

AlGhanim, M.A., Almohaimeed, S., Zheng, M., Solihin, Y. & Lou, Q. (2024). Jailbreaking LLMs with Arabic Transliteration and Arabizi. Available at: https://doi.org/10.48550/arXiv.2406.18725.

Bassignana, E., Basile, V. & Patti, V. (2018). Hurtlex: A multilingual lexicon of words to hurt. In *CEUR Workshop proceedings, Volume 2253*, pp. 1–6. CEUR-WS.

Cheng, M., Durmus, E. & Jurafsky, D. (2023). Marked personas: Using natural language prompts to measure stereotypes in language models. Available at: https://doi.org/10.48550/arXiv.2305.18189.

Chao, P., Robey, A., Dobriban, E., Hassani, H., Pappas, G. J. & Wong, E. (2023). Jailbreaking black box large language models in twenty queries. Available at: https://doi.org/10.48550/arXiv.2310.08419.

Choo, Y.H.M. & Bond, F. (2021). Taboo Wordnet. In P. Vossen & C. Fellbaum (eds.) *Proceedings of the 11th Global Wordnet Conference*. Potchefstroom: Global Wordnet Association, pp. 36–43.

Guzzetti, M. (2023). Forbidden Words and Female Anatomy. Gender and Language Taboos in the Oxford English Dictionary. *Lea*, 12, pp. 137–156. Available at: https://doi.org/10.36253/lea-1824-484x-14254.

Deng, Y., Zhang, W., Pan, S. J. & Bing, L. (2023). Multilingual jailbreak challenges in large language models. Available at: https://doi.org/10.48550/arXiv.2310.06474.

Fafalios, P., Iosifidis, V., Ntoutsi, E. & Dietze, S. (2018). Tweetskb: A public and large-scale rdf corpus of annotated tweets. In *European Semantic Web Conference* (pp. 177–190). Cham: Springer International Publishing.

Gupta, M., Charankumar A., Kshitiz A., Eli P. & Lopamudra, P. (2023). From ChatGPT to ThreatGPT: Impact of Generative AI in Cybersecurity and Privacy. *IEEE Access*. Available at: https://doi.org/10.1109/ACCESS.2023.3300381.

Honnavalli, S., Parekh, A., Ou, L., Groenwold, S., Levy, S., Ordonez, V. & Wang, W. Y. (2022). Towards understanding gender-seniority compound bias in natural language generation. Available at: https://doi.org/10.48550/arXiv.2205.09830.

Lazić, D. & Mihaljević, A. (2021). Social Stereotypes in Croatian Dictionaries from a Diachronic and a Synchronic Perspective. *Rasprave: Časopis Instituta za hrvatski jezik I jezikoslovlje*, 47(2), pp. 541–582.

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

Lewandowska-Tomaszczyk, B., Žitnik, S., Bączkowska, A., Liebeskind, C., Mitrović, J. & Valūnaitė Oleškevičienė, G. (2021). LOD-connected offensive language

ontology and tagset enrichment. In *CEUR workshop proceedings, Volume 3064*. Aachen: CEUR-WS.org.

Li, Z., Cabello, L., Yong, C. & Hershcovich, D. (2023). Cross-Cultural Transfer Learning for Chinese Offensive Language Detection. Available at: https://doi.org/ arXiv:2303.17927v1 [cs.CL]. (24 May 2023)

Merx, R., Vylomova, E. & Kurniawan, K. (2024). Generating bilingual example sentences with large language models as lexicography assistants. Available at: https://doi.org/10.48550/arXiv.2410.03182.

Naik, D., Naik, I. & Naik, N. (2024). Sorry, I am an AI language model: understanding the limitations of ChatGPT. In *The International Conference on Computing, Communication, Cybersecurity & AI*. Cham: Springer Nature Switzerland, pp. 26–42.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J. & Lowe, R. (2022). Training language models to follow instructions with human feedback. Available at: https://doi.org/10.48550/arXiv.2203.02155.

Phoodai, C., Rikk, R., Medveď, M., Měchura, M., Kosem, I., Kallas, J. & Jakubíček, M. (2023). Exploring the Capabilities of ChatGPT for Lexicographical Purposes: A Comparison with Oxford Advanced Learner's Dictionary within the Microstructural Framework. In G.B. van Huyssteen, C. Tiberius, M. Medveď, M. Měchura, I. Kosem, J. Kallas & S. Krek (eds.) *Electronic lexicography in the 21st century (eLex2023): Invisible Lexicography. Proceedings of the eLex2023 conference.* Brno: Lexical Computing CZ s.r.o., pp. 345–375.

Pronoza, E., Panicheva, P., Koltsova, O. & Rosso, P. (2021). Detecting ethnicity-targeted hate speech in Russian social media texts. *Information Processing & Management*, 58(6), 102674.

Qi, X., Zeng, Y., Xie, T., Chen, P. Y., Jia, R., Mittal, P. & Henderson, P. (2023). Fine-tuning aligned language models compromises safety, even when users do not intend to! Available at: https://doi.org/10.48550/arXiv.2310.03693.

Ruitenbeek, W., Zwart, V., Van Der Noord, R., Gnezdilov, Z. & Caselli, T. (2022). "Zo Grof !": A Comprehensive Corpus for Offensive and Abusive Language in Dutch. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH) Seattle, Washington.*

Shen, X., Chen, Z., Backes, M., Shen, Y. & Zhang, Y. (2024). "Do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pp. 1671–1685.

Tóth, Á. (forthcoming). TPEX: Neurális nyelvi modellek alkalmazása példamondatok kiválasztásában ['TPEX: The application of neural language models in selecting example sentences'].

Van Huyssteen, G.B. & Tiberius, C. (2023). Towards a lexical database of Dutch taboo language. 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. Brno: Lexical Computing CZ s.r.o., pp. 53–74.

Venkit, P.N., Gautam, S., Panchanadikar, R., Huang, T.H. & Wilson, S. (2023). Unmasking Nationality Bias: A Study of Human Perception of Nationalities in AI-Generated Articles. In F. Rossi, S. Das, J. Davis, K. Firth-Butterfield & A. John (eds.) *AIES 2023 - Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. New York: Association for Computing Machinery, pp. 554–565.

Xu, Z., Liu, Y., Deng, G., Li, Y. & Picek, S. (2024). A comprehensive study of jailbreak attack versus defense for large language models. Available at: https://doi.org/10.48550/arXiv.2402.13457.

Zaid, A. & Bennoudi, H. (2023). AI vs. Human Translators: Navigating the Complex World of Religious Texts and Cultural Sensitivity. *International Journal of Linguistics, Literature and Translation*, 6(11), pp. 173–182.

Identifying the Most Representative Phraseological Units Using Language Corpora and Artificial Intelligence for Lexicography: The Case of Slovenian Comparative Phrasemes

Matej Meterc, Nataša Jakop

ZRC, Fran Ramovš Institute of the Slovenian Language, Novi trg 2, SI-1000 Ljubljana, Slovenia

E-mail: matej.meterc@zrc-sazu.si, natasa.jakop@zrc-sazu.si

Abstract

In preparing phraseological units for the third edition of the *Standard Slovenian Dictionary* (eSSKJ), the authors aimed to identify the most relevant comparative phrasemes in the contemporary standard language using objective corpus-based criteria. A key goal was to determine how representative specific phrasemes and their variants are in actual use. Two lists of the hundred most frequent comparative phrasemes with the structure adjective + *kot* 'as' + noun (e.g., *bel kot sneg* 'white as snow') were extracted from the metaFida v1.0 corpus and CLASSLA-web.sl 1.0 corpora. The twenty most frequent were analyzed in greater detail. The results were compared with the Database of Comparative Phrasemes compiled from older dictionaries and collections, as well as with entries in eSSKJ. Artificial intelligence was also used experimentally to identify representative comparative phrasemes, with up to 80% alignment with expert choices.

Keywords: comparative phrasemes; corpus linguistics; artificial intelligence; lexicography; phraseological minimum

References

Bojc, E. (1987). *Pregovori in reki na Slovenskem*. Ljubljana: Državna založba Slovenije.

Čermák, F. (2007). *Frazeologie a idiomatika česká a obecná*. Prague: Univerzita Karlova v Praze, Karolinum.

Čermák, F. (2013). *Základní slovník českých přísloví*. Prague: Lidové noviny.

De Schryver, G.-M. (2023). Generative AI and Lexicography: The Current State of the Art Using ChatGPT. *International Journal of Lexicography* 36(4), pp. 355–387.

Dobrovol'skij, D. (2014). The Use of Corpora in Bilingual Phraseography. In A. Abel, C. Vettori & N. Ralli (eds.) *Proceedings of the XVI EURALEX International Congress. The User in Focus*. Bolzano: Institute for Specialised Communication and Multilingualism, pp. 867–885.

Ďurčo, P. (2014). Empirical Research and Paremiological Minimum. In H. Hrisztova-Gotthardt & M.A. Varga (eds.) *Introduction to Paremiology: A Comprehensive Guide to Proverb Studies*. Warsaw: Versita, pp. 183–205.

Erjavec, T. (2023). Corpus of Combined Slovenian Corpora metaFida 1.0, Slovenian Language Resource Repository CLARIN.SI, ISSN 2820-4042. Available at: http://hdl.handle.net/11356/1775.

eSSKJ: Slovar slovenskega knjižnega jezika. Accessed at: www.fran.si. (July 2025)

Fink-Arsovski, Ž., Kržišnik, E., Ribarova, S., Dunkova, T., Kabanova, N., Trostinska, R., Mironova Blažina, I., Spagińska-Pruszak, A., Vidović Bolt, I., Sesar, D., Dobríková, M. & Kursar, M. (2006). *Hrvatsko-slavenski rječnik poredbenih frazema*. Zagreb: Knjigra.

Gantar, P. (2002). Temeljne prvine zasnove frazeološkega slovarja. *Slavistična revija* 50(1), pp. 29–49.

Gantar, P. (2006). Corpus Approach in Phraseology and Dictionary Applications. *Slavistična revija* 54(1), pp. 161–162.

Gantar, P. (2007). *Stalne besedne zveze v slovenščini. Korpusni pristop.* Ljubljana: Založba ZRC, ZRC SAZU.

Hupkes, D., Dankers, V., Mul, M. & Bruniet, E. (2020). Compositionality Decomposed: How Do Neural Networks Generalise? *Journal of Artificial Intelligence Research* 67, pp. 757–795.

Jakubíček, M. & Rundell, M. (2023). The End of Lexicography? Can ChatGPT Outperform Current Tools for Post-Editing Lexicography? In: *Electronic Lexicography in the 21st Century (eLex 2023). Proceedings of the eLex 2023* Conference. Brno: Lexical Computing CZ, pp. 518–533.

Keber, J. (2011, 2015). *Slovar slovenskih frazemov*. Ljubljana: Založba ZRC, ZRC SAZU.

Kocijan, K., & Librenjak, S. (2016). Comparative Idioms in Croatian: MWU Approach. In G. Corpas Pastor (ed.) *Computerised and Corpus-Based Approaches to Phraseology: Monolingual and Multilingual Perspectives*. Geneva: Editions Tradulex, pp. 523–532.

Ljubešić, N., Rupnik, P. & Kuzman, T. (2024). *Slovenian Web Corpus CLASSLA-web.sl 1.0,* Slovenian Language Resource Repository CLARIN.SI, ISSN 2820-4042. Available at: http://hdl.handle.net/11356/1882.

Logar, N., Erjavec, T., Krek, S., Grčar, M., & Holozan, P. (2013). *Written Corpus ccGigafida 1.0*, Slovenian Language Resource Repository CLARIN.SI, ISSN 2820-4042. Available at: http://hdl.handle.net/11356/1035.

Meterc, M., & Jakop, N. (2016). Lexikografické spracovanie frazeologických variantov v novom slovníku slovinského spisovného jazyka. In M. Lišková (ed.) *Akademický slovník současné češtiny a software pro jeho tvorbu aneb Slovníky a jejich uživatelé v 21. století: sborník abstraktů z workshopu, Praha, 29.–30. listopadu 2016.* Prague: Ústav pro jazyk český AV ČR, pp. 55–56.

Meterc, M. (2017). *Paremiološki optimum: najbolj poznani in pogosti pregovori ter sorodne paremije v slovenščini.* Ljubljana: Založba ZRC, ZRC SAZU.

Meterc, M. (2020 –). *Slovar pregovorov in sorodnih paremioloških izrazov.* Available at: www.fran.si.

Meterc, M. & Mrvič, R. (in press). The Best-Known and Most Frequent Slovenian Proverbs, Listed by ChatGPT-4o: The Possibility to Create an AI-Supported/Based Paremiological minimum. *Linguistica.*

OpenAI (2025). *ChatGPT-4o (version from May 2024) [Large Language Model].* Accessed at: https://chat.openai.com. (9 July 2025).

Permjakov, G.L. (1971). *Paremiologicheskiy eksperiment: materialy dlya paremiologicheskogo minimuma.* Moscow: Nauka.

Permjakov, G. (1985). *300 obshcheupotrebitel'nykh russkikh poslovits i pogovorok (dlya govoryashchikh na nemetskom yazyke).* Moscow: Russkiy yazyk.

Varga, M.A., & Babić, S. (2023). Kroatische Sprichwortvarianten bei der Erstellung des kroatischen parömiologischen Thesaurus. *Yearbook of Phraseology* 14(1), pp. 147–164.

Wendler, C., Veselovski, V., Monea, G., & West, R. (2024). Do Llamas Work in English? On the Latent Language of Multilingual Transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* 1. Konstanz: KOPS Universität Konstanz, pp. 15366–15394.

An Electronic Ukrainian Dictionary as a Derussification Tool

Vasyl Starko[1], Andriy Rysin[2]

[1] Ukrainian Catholic University, 2a Kozelnytska Str., 79026 Lviv, Ukraine
[2] Independent researcher, Cary, NC, USA
E-mail: v.starko@ucu.edu.ua, arysin@gmail.com

Abstract

Due to the policy of Russification in the 20th century, the Ukrainian language underwent an influx of Russianisms, among other forms of interference with its structure. Today, many Ukrainians require guidance regarding non-Russified usage, and a *Large Electronic Dictionary of Ukrainian* (VESUM, vesum.nlp.net.ua) is designed to meet this need. With a register of over 430,000 lemmas, it is the most comprehensive morphological dictionary of Ukrainian. VESUM contains over 9,300 Russianisms, listed alongside their non-Russified equivalents. The decisions on what counts as a Russified item in need of replacement are based on multiple reputable sources, including dictionaries on the r2u.org.ua dictionary portal.

VESUM is the centerpiece of Pravopysnyk, the Ukrainian module of the LanguageTool text checker (check.nlp.net.ua, languagetool.org/uk). The role of VESUM is threefold. First, it supplies single-word Russified items and their replacements. Second, as a machine-readable dictionary, it serves as the source of data for lemmatization and morphological tagging, which are necessary for advanced text checking. Finally, VESUM can also be consulted as a stand-alone online dictionary via a web interface with flexible search options. As part of the Pravopysnyk tool, this electronic dictionary provides users with guidance on derussification when and where such advice is needed.

Keywords: electronic dictionary; Ukrainian; derussification; Russianism; error correction

References

BRUK: *Ukrainian Brown Corpus*. Accessed at: https://github.com/brown-uk/dict_uk. (7 July 2025)

Horodens'ka, K. (2019). *Ukrajins'ke slovo u vymirax s'ohodennja*. 2nd ed. Kyiv: KMM.

JMH: Antonenko-Davydovyč, B. *Jak my hovorymo*. Accessed at: http://yak-my-hovorymo.wikidot.com. (7 July 2025)

Karavans'kyj, S. (2001). *Pošuk ukrajins'koho slova, abo borot'ba za nacionalne "Ja."* Kyiv: Akademija.

Karavans'kyj, S. (2009). *Sekrety ukrajins'koji movy*. 2nd ed. L'viv: BaK.

KM: *Kultura movy na ščoden'*. Accessed at: http://kulturamovy.univ.kiev.ua. (7 July 2025)

Kryts'ka, V., Nedozym, T., Orlova, L., Puzdyrjeva, T., Romaniuk, Ju. (2011). *Hramatyčnyi slovnyk ukrajinskoji literaturnoji movy. Slovozmina*. Kyiv: Dmytro Burago Publishing House.

LANG-UK: Lang-uk. Accessed at: https://lang.org.ua/en/. (7 July 2025)

Masenko, L., Kubajčuk, V. & Dems'ka-Kul'čycka, O. (eds.) (2005). *Ukrajins'ka mova u XX storiččі: istorija linhvocydu: Dokumenty i materialy*. Kyiv: Kyjevo-Mohyljans'ka akademija.

NLP-UK: NLP-UK toolkit for Ukrainian. Accessed at: https://github.com/brown-uk/nlp_uk. (7 July 2025)

Ponomariv, O. (2011). *Kultura slova: Movnostylistyčni porady*. 4th ed. Kyiv: Lybid'.

Ponomariv, O. (2017). *Ukrajins'ke slovo dlja vsix i dlja kožnoho*. 2nd ed. Kyiv: Lybid'.

Pravopysnyk: Pravopysnyk LanguageTool. Available at: languagetool.org/uk, check.nlp.net.ua. (7 July 2025)

PULS: *Ukrainian Vocabulary Profile*. (2025). Lviv. Accessed at: https://puls.peremova.org. (7 July 2025)

R2U: Russian-Ukrainian dictionary portal. Accessed at: https://r2u.org.ua. (7 July 2025)

R2UF: R2U Forums. Accessed at: https://r2u.org.ua/forum. (7 July 2025)

RUS-1924: *Rosijs'ko-ukrajins'kyj slovnyk*. (1924-1933). Vols. 1–3, Kyiv, Xarkiv: DVU, URE.

RUS-2011: *Rosijs'ko-ukrajins'kyj slovnyk*. (2011-2014). Vols. 1–4, Kyiv: Znannja.

Serbens'ka, O. (2022). *Antysuržyk*. Lviv: Apriori.

Shevelov, G. (1966). Puryzm v ukrajinskij movi. *Ukrajinski visti,* 23, pp. 2–3. Available at: https://zbruc.eu/node/104522.

Shevelov, G. Y. (1989). *The Ukrainian Language in the First Half of the Twentieth Century (1900-1941): Its State and Status*. Cambridge, Mass.: Harvard University Press.

Shvedova, M. (2020). The General Regionally Annotated Corpus of Ukrainian (GRAC, uacorpus.org): Architecture and Functionality. In *Computational Linguistics and Intelligent Systems. Proc. 4th Intl. Conf. COLINS 2020*, pp. 489–506.

Shvedova, M., Waldenfels, R. von, Yarygin, S., Rysin, A., Starko, V., Nikolajenko, T. et al. (2017–2025). *GRAC: General Regionally Annotated Corpus of Ukrainian*. Kyiv, Lviv, Jena. Accessed at: http://uacorpus.org. (7 July 2025)

Starko, V. (2017). Kompjuterni linhvistyčni proekty hurtu r2u: stan ta zastosuvannja. *Ukrajins'ka mova*, 3, pp. 86–100.

Starko, V. & Rysin, A. (2020). Velykyj elektronnyj slovnyk ukrajinskoji movy (VESUM) jak zasib NLP dlja ukrajinskoji movy. In *Halaktyka Slova. Halyni Makarivni Hnatiuk*. Kyiv: Dmytro Burago Publishing House, Kyiv, pp. 135–141.

Starko, V. & Rysin, A. (2022). VESUM: A Large Morphological Dictionary of Ukrainian As a Dynamic Tool. *Computational Linguistics and Intelligent Systems. Proc. 6th Int. Conf. COLINS 2022*, Gliwice, Poland, pp. 71–80.

Starko, V. & Rysin, A. (2023). Creating a POS Gold Standard Corpus of Modern Ukrainian. *Proceedings of the Second Ukrainian Natural Language Processing Workshop* (UNLP). ACL, pp. 91–95.

UFLTC: *The UFL Textbook Corpus*. Accessed at: http://corpus-puls.peremova.org. (7 July 2025)

UP: *Ukrajins'kyj pravopys*. Accessed at: https://www.inmo.org.ua/pravopys-2019.html. (17 July 2025)

Vakulenko, S. (ed.) (2018). *Ukrajins'ka mova: unormuvannja, rozunormuvannja, perevnormuvannja* (1920-2015). Xarkiv: Xarkivs'ke istoryko-filolohične tovarystvo.

VESUM: Rysin, A. & Starko, V. *Large Electronic Dictionary of Ukrainian (VESUM)*. (2005-2025). Web version 6.7.1-SNAPSHOT. Accessed at: https://vesum.nlp.net.ua. (7 October 2025)

VESUM-GIT: Rysin, A. & Starko, V. *Large Electronic Dictionary of Ukrainian (VESUM)*. (2005-2025). Version 6.7.1-SNAPSHOT. Accessed at: https://github.com/brown-uk/dict_uk. (7 October 2025)

Vyxovanec', I. & Horodens'ka, K. (2004). *Teoretyčna morfolohijia ukrajins'koji movy*. Kyiv: Pul'sary.

GramatiKat: A Corpus-Based Tool for Detecting Morphological Anomalies and Paradigm Variation

Dominika Kovarikova

Institute of the Czech National Corpus, Charles University
E-mail: dominika.kovarikova@ff.cuni.cz

Abstract

GramatiKat is a freely accessible online application designed to support lexicographic and grammatical work on morphologically rich languages. It provides grammatical profiles, a frequency distribution of lemmas inflected forms, for thousands of Czech nouns, adjectives, and verbs based on large annotated corpora. The concept of grammatical profiling is rooted in the work of Janda and Lyashevskaya (2011), who demonstrated that the distribution of inflected forms can reflect both grammatical structure and semantic properties of lexemes. In *GramatiKat*, these profiles are compared against a statistically computed *Reference Grammatical Profile* (RGP), which captures the expected distribution of forms for a given part of speech (Kováříková & Nikolaev, in preparation). This allows users to immediately see whether a given word follows the expected distributional pattern or deviates from it in meaningful ways. Such deviations can signal lexicographically relevant features such as semantic anomalies or collocational behaviour (e.g. participation in multi-word terms, idioms, or other multi-word units).

The information in GramatiKat is derived from two representative corpora of contemporary written Czech, SYN2015 and SYN2020 (each containing 100 million words). Deviations from the norm, i.e. forms that are unusually frequent, infrequent, or entirely missing, are automatically highlighted using standard boxplot methodology (Kováříková & Kovářík 2023). Such anomalies can point to a wide range of lexicographically relevant information, including semantic constraints, syntactic preference, or idiomatic usage, all of which are valuable both for dictionary authors and for their audiences, particularly language learners.

The value of the tool for lexicographers is twofold. First, it offers empirical support for deciding whether certain grammatical forms should be included, exemplified, or specially marked in a dictionary entry. For instance, the noun *brva* 'eyelash' appears almost exclusively in the instrumental singular, as part of the idiom *nepohnout ani brvou* ('not to bat an eyelash'), which suggests that it is effectively defective in other forms (Kováříková et al. 2024), which is an information that should be included in the dictionary. Second, even when no overt anomaly is present, the grammatical profile provides a reliable picture of how a word behaves in real usage, for example showing the grammatical roles (nominative for subject, accusative for object). This supports

more nuanced dictionary descriptions in line with corpus-driven approaches that aim to derive linguistic generalizations directly from data (Tognini-Bonelli 2001).

From a technical perspective, *GramatiKat* lowers the barrier to corpus-based grammatical analysis by offering fully preprocessed, transparent, and reproducible data visualizations. The interface supports interactive exploration, filtering, and data export, making it accessible even to those without programming skills. The tool has already been successfully adapted to Slovak and Croatian, demonstrating that, given sufficient high-quality corpus data, the approach is transferable to other morphologically rich languages. Its development is grounded in principles of Open Science and reproducible research (Chromý & Cvrček 2021).

By combining grammatical profiling with robust statistical interpretation, *GramatiKat* equips lexicographers with a precise and efficient method for exploring morphological behavior across the lexicon. The presentation will illustrate the tool's functionality through real-world examples, showing both regular and anomalous grammatical profiles, and discussing how these can inform dictionary writing, editing, and revision.

Keywords: Tools for Lexicography; Grammatical Profiling; Morphological Anomalies;

Collocability; Corpus Analysis

1. References

Chromý, J. & Cvrček, V. (2021). Lingvistika jako otevřená a transparentní disciplína. *Naše řeč*, 104(4), pp. 233–243.

Janda, Laura A., & Lyashevskaya, O. (2011). Grammatical profiles and the interaction of the lexicon with aspect, tense, and mood in Russian. *Cognitive Linguistics* 22 (4), pp. 719–763.

Kováříková, D. & Kovářík, O. (2023). *GramatiKat: Online Tool for Grammatical Profiling*. Czech National Corpus. Accessed at: https://www.korpus.cz/gramatikat.

Kováříková, D. & Nikolaev, A. (in preparation). Methodological Considerations of Defectivity and Overabundance Analysis. [Manuscript under review]

Kováříková, D. et al. (2019). Lexicographer's Lacunas: Or How to Deal with Missing Representative Dictionary Forms on the Example of Czech. *International Journal of Lexicography*, 33(3), pp. 90–103.

Křen, M. et al. (2015). *SYN2015: Reprezentativní korpus psané češtiny*. Institute of the Czech National Corpus, Faculty of Arts, Charles University. Accessed at: https://www.korpus.cz.

Křen, M. et al. (2020). *SYN2020: Reprezentativní korpus psané češtiny*. Institute of the Czech National Corpus, Faculty of Arts, Charles University. Accessed at: https://www.korpus.cz.

Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work*. Amsterdam: John Benjamins.

Automated Transcription of Mixed-Script Dialectal Materials

Markus Kunzmann

Austrian Centre for Digital Humanities (ACDH), Austrian Academy of Sciences (ÖAW),
Bäckerstraße 13, 1010 Wien, Austria
E-mail: markus.kunzmann@oeaw.ac.at

Abstract

The project Dictionary of Bavarian Dialects in Austria "Wörterbuch der bairischen Mundarten in Österreich"(WBÖ) project maintains an archive of approximately 3.6 million handwritten dialectal paper slips documenting dialectal evidence. While 2.4 million entries have been manually digitized and converted to TEI format, the remaining 1.2 million paper slips from sections A-C require automated processing. This paper presents a novel three-stage workflow concept combining Handwritten Text Recognition (HTR) technology with existing digitized holdings to overcome the challenges posed by heterogeneous writing systems, multiple scribes, and poor material condition. Initial tests with existing HTR models yielded unsatisfactory results. The proposed solution leverages the existing Database of Bavarian Dialects "Datenbank der bairischen Mundarten in Österreich" (DBÖ) to automatically correct HTR transcription errors through similarity-based alignment and N-gram matching algorithms. The corrected transcriptions serve as a gold standard or a kind of ground truth for training a specialized HTR model tailored to historical dialect materials. This methodology enables the creation of substantial training datasets without manual transcription, potentially generating 33.6 million words for model training. The approach promises complete digital access to the WBÖ archive and provides a transferable template for similar lexicographic projects with historical slip collections.

Keywords: handwritten text recognition; dialect lexicography; digital humanities;

historical paper slips; workflow proposal

References

Barabas, B., Hareter-Kroiss, C., Hofstetter, B., Mayer, L., Piringer, B. & Schwaiger, S. (2010). Digitalisierung handschriftlicher Mundartbelege. Herausforderungen einer Datenbank. In *Germanistische Linguistik 199–201. Fokus Dialekt. Festschrift für Ingeborg Geyer zum 60. Geburtstag*, pp. 47–60.

Bauer, W. & Kühn, E. (1998). Vom Zettelkatalog zur Datenbank. Neue Wege der

Datenverwaltung und Datenbearbeitung im „Wörterbuch der bairischen Mundarten in Österreich". In C.J. Hutterer & G. Pauritsch (eds.) *Beiträge zur Dialektologie des ostoberdeutschen Raumes. Referate der 6. Arbeitstagung für bayerisch-österreichische Dialektologie, 20.–24.9.1995 in Graz*, number 636 in Göppinger Arbeiten zur Germanistik. Göppingen: Kümmerle Verlag, pp. 369–382.

Bowers, J. & Stöckle, P. (2018). TEI and Bavarian dialect resources in Austria: updates from the DBÖ and WBÖ. In A.U. Frank, C. Ivanovic, F. Mambrini, M. Passarotti & C. Sporleder (eds.) *Proceedings of the Second Workshop on Corpus-Based Research in the Humanities (CRH-2)*, volume 1 of *Gerastree proceedings*. Wien, pp. 45–54.

eScriptorium Project (2025). *Import data into eScriptorium*. École Pratique des Hautes Études, AOROC. Available at: https://escriptorium.readthedocs.io/en/latest/import/. User documentation for data import in eScriptorium HTR platform.

Frank, A.U., Ivanovic, C., Mambrini, F., Passarotti, M. & Sporleder, C. (eds.) (2018). *Proceedings of the Second Workshop on Corpus- Based Research in the Humanities CRH-2*, volume 1 of *Gerastree proceedings*.

Hornung, M. & Bauer, W. (1983). *Wörterbuch der bairischen Mundarten in Österreich (WBÖ). Band 3: Pf – C*, volume 3. Wien: Verlag der Österreichischen Akademie der Wissenschaften.

Kahle, P., Colutto, S., Hackl, G. & Mühlberger, G. (2017). Transkribus - A Service Platform for Transcription, Recognition and Retrieval of Historical Documents. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 04, pp. 19–24.

Kranzmayer, E. (1970). *Wörterbuch der bairischen Mundarten in Österreich (WBÖ). Band 1: A – Azor*, volume 1. Wien: Kommissionsverlag der Österreichischen Akademie der Wissenschaften.

Kranzmayer, E. (1976). *Wörterbuch der bairischen Mundarten in Österreich (WBÖ). Band 2: B(P) – Bezirk*, volume 2. Wien: Verlag der Österreichischen Akademie der Wissenschaften. Maria Hornung (Redaktion).

Lenz, A.N. & Stöckle, P. (eds.) (2021). *Germanistische Dialektlexikographie zu Beginn des 21. Jahrhunderts*. Stuttgart: Steiner. Unter Mitarbeit von A. Bergermayer, A. Gellan, S. Wahl, E. Wahlmüller & P. Zeitlhuber.

Puigcerver, J. (2017). Are Multidimensional Recurrent Layers Really Necessary for Handwritten Text Recognition? In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 01, pp. 67–72.

READ-Coop. (2022). The road to Handwritten Text Recognition – Part 1. Accessed at: https://blog.transkribus.org/en/insights/road-to-htr-1. (30 September 2025)

READ-Coop. (2023). Introducing Transkribus Super Models – Get access to the Text Titan I. Accessed at: https://blog.transkribus.org/en/introducing-transkribus-super-models-get-access-to-the-text-titan-i. (30 September 2025)

Stöckle, P. (2021). Wörterbuch der bairischen Mundarten in Österreich (WBÖ). In A.N. Lenz & P. Stöckle (eds.) *Germanistische Dialektlexikographie zu Beginn des*

21. Jahrhunderts. Stuttgart: Steiner, pp. 11–46. Unter Mitarbeit von A. Bergermayer, A. Gellan, S. Wahl, E. Wahlmüller & P. Zeitlhuber. Transkribus (2025). Data Preparation. Available at: https://help.transkribus.org/data-preparation.

Wiesinger, P. (1964). Das Phonetische Transkriptionssystem Der Zeitschrift 'Teuthonista'. Eine Studie Zu Seiner Entstehung Und Anwendbarkeit in Der Deutschen Dialektologie Mit Einem Überblick Über Die Geschichte Der Phonetischen Transkription Im Deutschen Bis 1924. *Zeitschrift für Mundartforschung*, 31(1), pp. 1–20. Available at: http://www.jstor.org/stable/40500597.

Why a dedicated dictionary device is more appropriate than an app for primary school learners

Lorna Morris

Stellenbosch University
E-mail: lorna@lemma.co.za

Abstract

South Africa is in a literacy crisis, with learners not progressing in school because they are being taught in a second language when they are not functionally literate in their first language. Fewer than 10% of South Africans have English as a home language, but 90% of learners are being taught in English. Many South African schools are under resourced and are not able to give learners the support they need. An e-dictionary has been designed to combat literacy amongst primary school learners. This dictionary contains audio for the pronunciation of the headword, meaning, and examples; hyperlinks connect semantically related entries; full colour illustrations illustrate every sense of every word; and home language translation equivalents of the headword are presented at each sense. These are some of the features that provide extra support for learners learning in their second language.

In terms of the medium on which to supply an e-dictionary to learners, there are three options: an online dictionary accessible to anyone with a device and internet access; an app that is accessible to anyone with a smart phone or tablet; and a dedicated dictionary device that does not require electricity or access to the internet. Many people suggest that since almost all adults are in possession of a smart phone, an app would be the most obvious solution. This paper shows that for South African primary school learners living under the circumstances described above, a dedicated dictionary device is the better option. This conclusion is based on research that has been done in under resourced primary schools in three provinces in South Africa. This research comprised of classroom observations of Grade 5 and 6 learners using a model dictionary on a stand-in device; focus group discussions with learners who had been using these devices; interviews with class and language teachers; and interviews with South African literacy experts. The reasons given for the preference for a device over an app include firstly, that it minimises distractions typically associated with smart phones and tablets, such as a camera and other apps. The device would need to be cost-effective, addressing the financial constraints faced by most South African schools, and it would need to be more robust than smart phones and tablets, to ensure durability in diverse and often challenging environments. These reasons were echoed by learners, teachers, and literacy experts. The paper will present the results of the research and show why a dedicated dictionary device is more suitable than an app for primary school learners.

Keywords: school dictionary; pedagogical lexicography; PED; portable dictionary device; electronic dictionary

Modeling and structuring of a bilingual French-Chinese phraseological dictionary: neural automatic approach for ontology and lexicography

Lian Chen 陈恋

LLL – University of Orleans, France
CRLAO-CNRS-INALCO, France
E-mail: lian.chen@univ-orleans.fr

Abstract

The creation of ontologies—traditionally the domain of linguists and knowledge engineers—is undergoing a significant transformation thanks to advances in artificial intelligence and natural language processing (NLP). These developments open new avenues for phraseology, a field where multi-word expressions (MWEs)—often opaque and non-compositional—must be identified, classified, and linked to abstract concepts or discourse contexts (Constant 2012: 6). Despite their linguistic richness, idiomatic expressions remain a major challenge for NLP due to their syntactic variability, semantic ambiguity, and context-dependence (Gross 1996; Mejri 1997; Polguère 2002; Chen 2021).

This study presents an approach for modeling a bilingual French–Chinese phraseological dictionary by combining lexicographic theory, ontology design, and neural NLP techniques. We focus specifically on idiomatic expressions related to the **human body** and **animals**, domains in which words such as *main* (hand) can carry both literal and figurative meanings—e.g., as symbols of work, strength, or authority (Rey & Chantreau 2003; Rey 2019).

To overcome the limitations of manual ontology construction tools like Protégé (Kapoor & Sharma, 2010), we follow the principles of the Ontology Layer Cake (Despres & Szulman 2008; Tiwari & Jain 2014) and implement a semi-automatic pipeline. Our methodology includes: (1) statistical extraction of idioms using TF-IDF, PMI, and RAKE; (2) syntactic filtering of candidate MWEs; (3) visualization and annotation through an interactive Streamlit interface; (4) semantic relation modeling using fine-tuned neural models (BilBERT and Sentence-BERT); and (5) export in OWL/RDF format using the OntoLex-Lemon standard, with SKOS for conceptual hierarchies and VarTrans for bilingual alignments.

A central challenge lies in extracting **semantic triplets** of the form *(idiom, keyword, relation)*—e.g., *donner un coup de main* → *(main, aide)*—which requires addressing the idioms' non-compositionality, structural variation, and semantic opacity. We rely

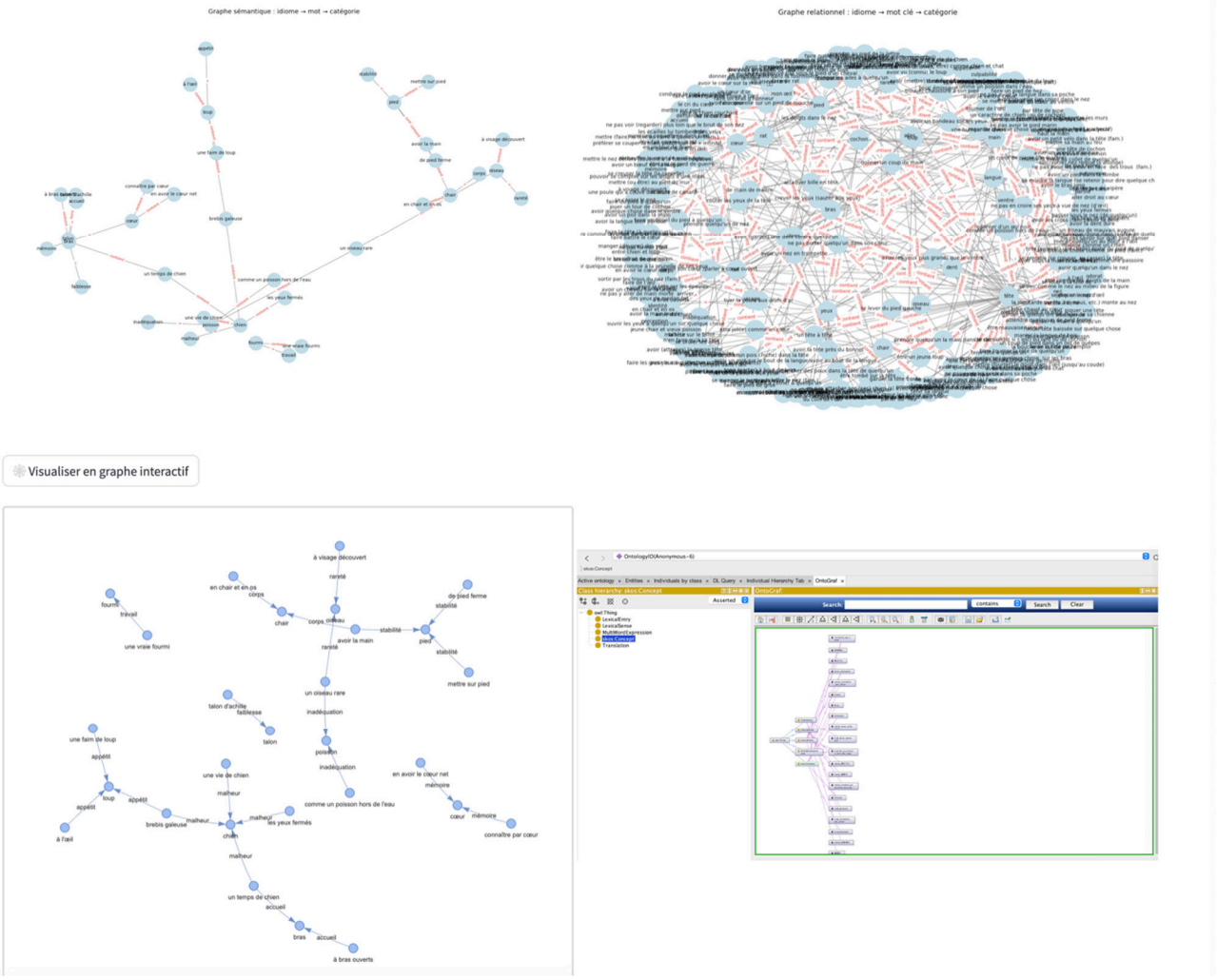
on syntactic grammars (Tesnière 1959), semantic mapping, and machine learning to formalize these triplets into interpretable ontological structures (Chen & Gasparini 2025).

The resulting resource is a multilingual, interoperable, and dynamic dictionary of idiomatic expressions, accessible via an interface that supports exploration, sorting, and export to Protégé or SPARQL-compatible systems. This work bridges NLP and lexicography, contributing to AI-enhanced **auto-lexicography**, semantic modeling, and the generation of context-aware bilingual examples (González-Rey 2002; Mel'čuk 2008, 2011; Mejri 2011; Sułkowska 2016; Chen 2023).

Our project aims to achieve six interconnected objectives. First, we design a semi-automatic pipeline for extracting and identifying idiomatic expressions from authentic French corpora, with a particular focus on thematic categories such as the human body and animals. Second, we construct semantic triplets that link idioms to keywords and conceptual categories, enabling fine-grained semantic interpretation. Third, we fine-tune a multilingual BERT-based model (BilBERT) to classify the semantic relations between idioms and their components. Fourth, we formally model the extracted data as an ontology using the OntoLex-Lemon framework, enriched with SKOS hierarchies and VarTrans modules to support bilingual alignment with Chinese equivalents. Fifth, we develop an interactive Streamlit interface that allows users to visualize idiomatic relationships, perform manual annotations, and export the data in RDF/OWL format. Finally, our project contributes to ongoing research in multilingual phraseology and AI-assisted lexicography, offering practical tools and resources for Semantic Web applications and advanced NLP tasks.

Here are several illustrations of the results obtained throughout the project, including visualizations of idiomatic triplets, conceptual mappings, and semantic graphs generated during the modeling and classification phases.

Keywords: auto-lexicography; ontological relations automation; knowledge engineering; natural language processing; e-lexicography



References

Amdouni, E., Belfadel, A., Gagnant, M., Renault, I., Kierszbaum, S., Carrion, J., Dussartre, M. & Tmar, S. (2025). Semi-Automatic Building of Ontologies from Unstructured French Texts: Industrial Case Study. *Data Science and Engineering*. Published 19 June 2025. Available at: https://doi.org/10.1007/s41019-025-00284-z.

Chen, L. & Gasparrini, N. (2025). Modélisation et structuration d'un dictionnaire bilingue français-chinois des expressions idiomatiques : approche lexicographique et ontologique. In *Proceedings of the Sixième Colloque international « Dictionnaire et polylexicalité »*, Université de Bari (Italie), Université Sorbonne Paris Nord (France), Université de Silésie à Katowice (Pologne).

Chen, L. (2024). Traitement de la traduction et de la transmission culturelle de la microstructure dans les dictionnaires bilingues des UP: étude et analyse contrastive de corpus métalexicographique. *SHS Web of Conferences*, 139, 11001, pp. 1–18.

Chen, L. (2023). (Meta)phraseography and phraseomatics: DiCoP, a computerized

resource of phraseological units. In *Conference Proceedings of ASIALEX 2023: Lexicography, Artificial Intelligence, and Dictionary Users – The 16th International Conference of the Asian Association for Lexicography*, pp. 224–231.

Chen, L. (2021). *Analyse comparative des expressions idiomatiques en chinois et en français (relatives au corps humain et aux animaux)* [PhD thesis, Cergy Paris Université].

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⟩

Despres, S. & Szulman, S. (2008). Réseau terminologique versus Ontologie. In *Proceedings of Toth 2008*, France, pp. 1–19.

Dessì, D., Osborne, F., Reforgiato Recupero, D., Buscaldi, D. & Motta, E. (2020). Generating Knowledge Graphs by Employing Natural Language Processing and Machine Learning Techniques within the Scholarly Domain. Available at: https://arxiv.org/pdf/2011.01103.

Elnagar, S., Yoon, V. & Thomas, M.A. (2020). An Automatic Ontology Generation Framework with An Organizational Perspective. In *Proceedings of the 53rd Hawaii International Conference on System Sciences*, pp. 4860–4869.

EUROPHRAS. (2023). 4th International Conference on Computational and Corpus-based Phraseology. In *Proceedings of EUROPHRAS 2023*, pp. 17–25.

González-Rey, M.I. (2002). *La phraséologie du français.* Toulouse: Presses Universitaires du Mirail.

Kapoor, B. & Sharma, S. (2010). A Comparative Study of Ontology Building Tools for Semantic Web Applications. *International Journal of Web & Semantic Technology (IJWesT)*, 1(3), pp. 1–13.

OLAF: An Ontology Learning Applied Framework. (2023). In *Proceedings of the 27th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems (KES 2023)*, Athens, Greece, pp. 2106–2115.

Mejri, S. (2011). Phraséologie et traduction. *Équivalence*, 38(1–2), pp. 111–133.

Mejri, S. (1997). *Le figement lexical: descriptions linguistiques et structuration sémantique* (Série Notions de base en lexicologie). Manouba: Publications de la Faculté des Lettres de la Manouba.

Mel'čuk, I. (2008). La phraséologie et son rôle dans l'enseignement/apprentissage d'une langue étrangère. *Études de Linguistique Appliquée*, 92, pp. 82–117.

Mel'čuk, I. (2011). Phrasèmes dans le dictionnaire. In J.-C. Anscombre & S. Mejri (eds.) *Le figement linguistique : la parole entravée*, Paris: Honoré Champion, pp. 41–61.

Mel'čuk, I. (2013). Tout ce que nous voulions savoir sur les phrasèmes, mais…. *Cahiers de Lexicologie*, 102, pp. 129–149.

Murano, M. (2011). *Le traitement des séquences figées dans les dictionnaires bilingues français-italien, italien-français.* [Édition en français].

Musen, M.A. (2015). The Protégé project: A look back and a look forward. *AI Matters*, 1(4), pp. 4–12.

Polguère, A. (2008). *Lexicologie et sémantique lexicale*. Montréal: Les Presses de l'Université de Montréal.

Polguère, A. (2002). *Notions de base en lexicologie*. Paris: Ophrys.

Sułkowska, M. (2016). Phraséodidactique et phraséotraduction : quelques remarques sur les nouvelles disciplines de la phraséologie appliquée. *Yearbook of Phraseology*, 7, pp. 35–54.

Tiwari, S.M. & Jain, S. (2014). Automatic Ontology Acquisition and Learning. *IJRET: International Journal of Research in Engineering and Technology*, pp. 38–43. eISSN: 2319-1163 | pISSN: 2321-7308.

Varone, M. (2011). Method and System for Automatically Extracting Relations Between Concepts Included in Electronic Text. *U.S. Patent* 7,899,666 B2, issued March 1, 2011. Assignee: Expert System S.p.A. (Modena, Italy). Application filed May 4, 2007.

Image-to-Sense Alignment Using AI Tools

Andrej Perdih, Dejan Gabrovšek, Janoš Ježovnik

ZRC SAZU, Fran Ramovš Institute of the Slovenian Language, Novi trg 2, 1000 Ljubljana, Slovenia

E-mail: andrej.perdih@zrc-sazu.si, dejan.gabrovsek@zrc-sazu.si, janos.jezovnik@zrc-sazu.si

Abstract

This paper evaluates the results of using GPT-4o mini language model batch processing with image recognition capability to align 1,572 images of 398 polysemous nouns in the Dictionary of the Slovenian Standard Language (second edition) to their specific dictionary senses, and it compares them to the results of the manual image-to-sense alignment process. The images were manually assigned to entries in a previous task, but no sense information was provided at the time. The language model showed relatively high overall agreement with the human annotator (i.e., 85.1%). In cases in which multiple senses were selected per image in both manual and automated annotation, the agreement was even slightly higher (i.e. in 89.4% of all sense evaluations). The agreement rate was higher when the language model evaluated only the matching senses and lower when it also evaluated the non-matching senses within the entry.

Keywords: images; lexicography; image-to-sense alignment; image recognition

References

Biesaga, M. (2016). Pictorial Illustration in Dictionaries The State of Theoretical Art. In T. Margalitadze & G. Meladze (eds.) *Proceedings of the 17th EURALEX International Congress, EURALEX 2016*. Tbilisi: Lexicographic Centre, Ivane Javakhishvili Tbilisi State University, pp. 99–108. Available at: http://euralex.org/wp-content/themes/euralex/proceedings/Euralex%202016/euralex_2016_007_p99.pdf.

Biesaga, M. (2017a). Pictorial Illustrations in Encyclopaedias and in Dictionaries – a Comparison. In I. Kosem, C. Tiberius, M. Jakubíček, J. Kallas, S. Krek, & V. Baisa (eds.) *Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference*. Lexical Computing CZ, pp. 221–236. Available at: https://elex.link/elex2017/wp-content/uploads/2017/09/paper13.pdf.

Biesaga, M. (2017b). Dictionary Tradition vs. Pictorial Corpora: Which Vocabulary Thematic Fields Should Be Illustrated? *Lexikos*, 27, pp. 132–151. Available at: https://doi.org/10.5788/27-1-1397.

De Schryver, G.-M. (2023). Generative AI and Lexicography: The Current State of the

Art Using ChatGPT. *International Journal of Lexicography*, 36(4), pp. 355–387. Available at: https://doi.org/10.1093/ijl/ecad021.

Dziemianko, A. (2022). The usefulness of graphic illustrations in online dictionaries. *ReCALL*, 34(2), pp. 218–234. Available at: https://doi.org/10.1017/S0958344021000264.

Dziemianko, A. (2024). Pictures in Online Dictionaries: Shall We See Them? *GEMA Online® Journal of Language Studies*, 24(2), pp. 31–53. Available at: https://doi.org/10.17576/gema-2024-2402-03.

Franček. Accessed at: https://www.franček.si/. (7 July 2025)

Kallas, J., Koppel, K., & Tsepelina, K. (2024). The EKI Picture Dictionary. A Multilingual Tool for A1–B1 Learners. In K. Š. Despot, A. Ostroški Anić, & I. Brač (eds.) *Lexicography and Semantics. Book of Abstracts of the XXXI EURALEX International Congress. Institut za hrvatski jezik*, pp. 163–165. Available at: https://euralex.jezik.hr/wp-content/uploads/2021/09/Euralex_boa_20.pdf.

Krek, S. (ed.). (2024). *Book of Abstracts of the Workshop Large Language Models and Lexicography*. [Pre-publication]. Available at: https://www.cjvt.si/wp-content/uploads/2024/10/LLM-Lex_2024_Book-of-Abstracts.pdf.

Lew, R. (2010). Multimodal Lexicography: The Representation of Meaning in Electronic Dictionaries. *Lexikos*, 20, pp. 290–306. Available at: https://doi.org/10.5788/20-0-144.

Lew, R., Kaźmierczak, R., Tomczak, E., & Leszkowicz, M. (2018). Competition of Definition and Pictorial Illustration for Dictionary Users' Attention: An Eye-Tracking Study. *International Journal of Lexicography*, 31(1), pp. 53–77. Available at: https://doi.org/10.1093/ijl/ecx002.

Lew, R. (2024). Dictionaries and lexicography in the AI era. *Humanities and Social Sciences Communications*, 11(1). Available at: https://doi.org/10.1057/s41599-024-02889-7.

Lišková, M., & Šemelík, M. (2024). Show me the meaning of being lonely... Graphic Illustrations in The Academic Dictionary of Contemporary Czech. In K. Š. Despot, A. Ostroški Anić, & I. Brač (eds.) *Lexicography and Semantics. Book of Abstracts of the XXXI EURALEX International Congress. Institut za hrvatski jezik*, pp. 163–165. Available at: https://euralex.jezik.hr/wp-content/uploads/2021/09/ Euralex_boa_20.pdf.

Liu, X. (2015). Multimodal Definition: The Multiplication of Meaning in Electronic Dictionaries. *Lexikos*, 25, pp. 210–232. Available at: https://doi.org/10.5788/25-1-1296.

Perdih, A., Ahačič, K., Jakop, N., Ledinek, N., & Petric Žižić, Š. (2024). Semantic Information on the Franček Educational Language Portal for Slovenian. In K. Š. Despot, A. Ostroški Anić, & I. Brač (eds.) *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress. Institut za hrvatski jezik*, pp. 155–168. Available at: https://euralex.jezik.hr/wp-content/uploads/2021/09/Euralax-XXI-final-web.pdf.

Perdih, A., Ahačič, K., Ježovnik, J., & Race, D. (2021). Building an Educational Language Portal Using Existing Dictionary Data. *Journal of Linguistics/Jazykovedný časopis*, 72(2), pp. 568–578. Available at: https://doi.org/10.2478/jazcas-2021-0052.

SSKJ2: *Slovar slovenskega knjižnega jezika*. (2014). 2nd edition. Ljubljana: Inštitut za slovenski jezik Frana Ramovša ZRC SAZU.

Matching meaning: Evaluating ChatGPT's ability to assign corpus examples to dictionary senses of polysemous sound-related verbs

Sylwia Wojciechowska

Faculty of English, Adam Mickiewicz University, Poznań
E-mail: sylwiab@amu.edu.pl

Abstract

A major change in dictionary exemplification was brought about by the arrival of corpus data, which replaced lexicographer-made examples with authentic ones from real spoken and written discourse. Monolingual English learners' dictionaries (MELDs) prefer a third type of examples, corpus-based ones, with unnecessarily complex vocab and structure, and unclear content removed from them. However, apart from corpus-based examples which follow definitions of senses, MELDs online include sections of non-modified corpus examples placed usually at the bottom of entries and not matched with any senses.

The paper aims to explore corpus examples sections accompanying polysemous sound-related verbs and leverage ChatGPT-4 to match corpus examples with the senses already distinguished in the respective dictionary entries. The verbs were selected from the twelve strongest and forty-four strong synonym matches of the verb 'sound' in the sense "produce noise" on Thesaurus.com. Apart from the basic, literal meaning, each of these verbs has a figurative, metaphorical meaning or meanings, e.g. echo "to repeat opinions in agreement", and resonate "to receive a sympathetic response". Learners' dictionaries were chosen for analysis, as exemplification is particularly important in them. The selected MELDs are Longman Dictionary of Contemporary English (LDOCE), Cambridge Advanced Learner's Dictionary (CALD) and Collins Dictionary (Collins), as they all have sections dedicated to corpus examples. CALD and Collins explicitly inform the user that the examples have been automatically selected, and therefore the editors do not take responsibility for possible sensitive content or mismatches with the entry word.

The present study demonstrates that ChatGPT is successful at separating literal from metaphorical examples of sound-related verbs, which is not surprising, as current research indicates the capability of Large Language Models (LLMs) for polysemy and metaphor identification and interpretation (e.g. Bond et al. 2024 and Lin et al. 2024). The performance of ChatGPT is then checked in a more challenging task, that of matching corpus examples with the already existing senses in each of the analysed dictionaries. The prompts include the numbered senses that feature in the dictionaries

under a certain headword together with the definitions and accompanying examples, which serve as models for ChatGPT.

The corpus examples sections in the dictionaries tend to be rather lengthy, especially in CALD, and, for instance, at the entry for 'resonate' they amount to 104 examples. Therefore, the task of assigning corpus examples to separate senses would be drudgery for human lexicographers. In online dictionaries, such corpus examples can be located below corpus-based examples in expandable boxes, a practice which is already seen in Oxford Advanced Learner's Dictionary for corpus-based examples. It was found that sometimes ChatGPT admits it cannot assign any corpus example to a sense, because no example demonstrates it. Such cases will be analysed with scrutiny, and ChatGPT will be asked to generate missing examples, a task which it does not turn out to be impressive at, as Lew (2023) observes.

Keywords: Monolingual English learners' dictionaries; Corpus examples; ChatGPT;

Polysemous verbs; Metaphor

References

Dictionaries:

Cambridge Advanced Learner's Dictionary (CALD). Accessed at: http://www.dictionary.cambridge.org. (4 April 2025)

Collins Dictionary (Collins). Accessed at: http://www.collinsdictionary.com. (4 April 2025)

Longman Dictionary of Contemporary English (LDOCE). Accessed at: http://www.ldoceonline.com. (4 April 2025)

Other references:

Bond, F., Maziarz, M., Piotrowski, T. & Rudnicka, E. (2024). "Models of Polysemy in Two English Dictionaries", *International Journal of Lexicography*, 37(2), pp. 196–225.

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

Lin, Y., Liu, J., Gao, Y., Wang, A. & Su, J. (2024). "A Dual-Perspective Metaphor Detection Framework Using Large Language Models". Available at: http://arxiv.org/abs/2412.17332. (10 March 2025)

Inductive Categorization for Conceptual Analysis with LLMs: A Case Study from the Humanitarian Encyclopedia

Loryn Isaacs, Santiago Chambó, Pilar León-Araúz

University of Granada, Department of Translation and Interpreting, Puentezuelas, 55, 18071, Granada, Spain Affiliation of Author1, Address
E-mail: lisaacs@ugr.es, santiagochambo@ugr.es, pleon@ugr.es

Abstract

Corpus-based conceptual analysis for the Humanitarian Encyclopedia (HE) grapples with vast amounts of lexical data to describe the meaning of key humanitarian notions and detect conceptual variation among actors (Odlum & Chambó, 2022). By building on Frame-based Terminology (Faber, 2015, 2022), the HE is incorporating qualitative methods necessary to subsume lexical data into manageable semantic triples in a way that ensures the traceability and transparency of modeling decisions.

While traditional inductive qualitative analysis is labor-intensive, researchers are now replicating these methods using LLM-assisted workflows. Following this trend, our paper presents an observational study with a dataset of 274 spans labeled as causes of forced displacement that were manually annotated on a random sample of 1,000 concordances obtained from an English corpus of humanitarian documents from ReliefWeb (Isaacs et al., 2024). In this initial assessment, we test LLM inductive categorization using four models locally: Magistral Small 1.0 (Mistral-AI et al., 2025) with 24 billion parameters and three DeepSeek R1 models (DeepSeek-AI, 2025), with 8, 32 and 70 billion parameters. They are evaluated against a manual categorization comprising 34 causality groupings produced by two annotators through consensus.

To assess baseline similarities, we provide models with minimal, zero-shot instruction, while also requiring structured outputs and conducting 40 runs per model (10 runs per text format: lines, CSV rows, JSON dictionary and Python list). We evaluate model fitness by measuring (1) degree of task completion, (2) category assignment similarity to the gold standard and (3) semantic overlap of LLM-generated category labels with those in the gold standard. For category assignment similarity, multiple Jaccard similarity scores were converted into a single normalized measure. Category labels from the top ten runs (those exhibiting the highest degree of category assignment similarity) demonstrated semantic overlap with manual labels. Nevertheless, the results were mixed: some LLM-generated labels were invalid, whereas others, although absent from the gold standard, were considered pertinent by the annotators.

In conclusion, models displayed low overall similarity scores when given little instruction and hundreds of spans to classify in one batch, consistently omitting spans

despite being prompted not to do so. Outlier runs achieved similarity scores comparable to annotators, while revealing useful insights not captured in the manual categorization. The results underscore the complexity of categorizing data for a single, domain-specific concept. However, this also highlights the potential of LLMs as complementary tools for qualitative analysis tasks in the conceptual analysis workflow of the HE. Future work will investigate multi-category tasks, hybrid human-in-the-loop approaches, refined prompting strategies, and additional pre- and post-processing of lexical data.

Keywords: inductive categorization; humanitarian domain; large language model; structured output; conceptual analysis

References

DeepSeek-AI. (2025). DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. Available at: https://arxiv.org/abs/2501.12948.

Faber, P. (2015). Frames as a framework for terminology. In H. J. Kockaert & F. Steurs (eds.) *Handbook of Terminology: Volume 1.* Amsterdam: John Benjamins, pp. 14-33. Available at: https://benjamins.com/catalog/hot.1.fra1.

Faber, P. (2022). Chapter 16. Frame-based Terminology. In P. Faber & M.-C. L'Homme (eds.) *Theoretical Perspectives on Terminology: Explaining terms, concepts and specialized knowledge.* Amsterdam: John Benjamins, pp. 353-376. Available at: https://doi.org/10.1075/tlrp.23.16fab.

Isaacs, L., Chambó, S. & León-Araúz, P. (2024). Humanitarian Corpora for English, French and Spanish. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (eds.) *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 8418–8426. Available at: https://aclanthology.org/2024.lrec-main.738.

Mistral-AI, Rastogi, A., Jiang, A. Q., Lo, A., Berrada, G., Lample, G., Rute, J., Barmentlo, J., Yadav, K., Khandelwal, K., Chandu, K. R., Blier, L., Saulnier, L., Dinot, M., Darrin, M., Gupta, N., Soletskyi, R., Vaze, S., Scao, T. L., … Tang, Y. (2025). Magistral. Available at: https://doi.org/10.48550/arXiv.2506.10910.

Odlum, A. & Chambó, S. (2022). *Horizontally integrating diverse definitions and debates on key concepts in an online Humanitarian Encyclopedia.* Accessed at: https://humanitarianencyclopedia.org/expertise-note/horizontally-integrating-diverse-definitions-and-debates-on-key-concepts-in-an-online-humanitarian-encyclopedia. (15 July 2025)

Navigating linguistic diversity: modelling diatopic and bibliographic information with TEI Lex-0

Veronika Engler[1], Karlheinz Mörth[1], Stephan Procházka[2],

Michaela Rausch-Supola[1], Daniel Schopper[1]

[1] Austrian Centre for Digital Humanities, Austrian Academy of Sciences, Bäckerstraße 13, 1010 Vienna, Austria

[2] University of Vienna, Department of Near Eastern Studies, Spitalgasse 2, Court 4.1, 1090 Vienna, Austria

E-mail: {veronika.engler, karlheinz.moerth, michaela.rausch-supola, daniel.schopper}@oeaw.ac.at, stephan.prochazka@univie.ac.at

Abstract

The *Vienna Corpus of Arabic Varieties* (VICAV) is a digital research infrastructure for the documentation and analysis of the linguistic diversity of Arabic varieties͡. Integrating methods from language technology and the digital humanities, VICAV provides a modular, sustainable platform for the creation, management, and publication of heterogeneous language resources within a shared data architecture (Budin et al. 2012; Moerth et al. 2015). At its core lies a commitment to openness, interoperability, and adherence to community standards, in particular the *Guidelines* of the Text Encoding Initiative (TEI Consortium 2025). Through a text-centered, standards-based design, VICAV enables the representation of diverse types of data—including an extensive bibliography, linguistic profiles, sample texts, and digital dictionaries—within a unified technical framework and a user-friendly web application (https://vicav.acdh.oeaw.ac.at).

Among VICAV's key components are dictionaries of four Arabic varieties—Baghdad, Cairo, Damascus, Tunis—next to a dictionary of Modern Standard Arabic which mainly serves as a point of reference for the others (Procházka & Moerth 2015). These compact lexical databases, containing up to 8,000 entries each, provide structured lexicographic information enriched with English translations and, in some cases, also German, French, or Spanish. All are built on a shared TEI-based model ensuring consistent encoding and comparability across varieties.

The newest addition to the VICAV family of lexicographic resources is the SHAWI Dictionary, developed within the SHAWI Project (*The Shawi-type Arabic dialects spoken in South-eastern Anatolia and the Middle Euphrates region*, FWF P-33574, 2021–2027). The project investigates the varieties spoken by Bedouin communities in Turkey, Syria, Lebanon, and Iraq—which so far received little systematic attention by linguistic research. These dialects display internal variation which shows significant

geographic and sociolinguistic distribution—dimensions that require fine-grained modelling beyond the capabilities of standard TEI constructs. The SHAWI Dictionary, scheduled for a beta release in late 2025, represents the first VICAV dictionary encoded entirely in TEI Lex-0, a refinement of the TEI Dictionary Module developed by the DARIAH Working Group on Lexical Resources which aims at harmonizing the representation of lexical data and facilitating interoperability across projects (Tasovac et al., 2018ff.).

The adoption of TEI Lex-0 allows for both greater formal consistency and project-specific adaptability. The SHAWI Dictionary extends Lex-0 through the TEI mechanism of ODD chaining (Rahtz 2014), producing a VICAV-wide *generic dictionary schema* that forms a common backbone for future resources. The SHAWI Dictionary's project-specific adaption of this schema introduces several innovations:

(1) Encoding structures for **diatopic and sociocultural variation**: The element `<usg type="geographic">` serves as a wrapper to embedded `<name>` elements for places and tribes alike which are further linked to entities in local reference resources established in the project *WIBARAB* (*What is Bedouin-Type Arabic?* 2021-2026; ERC 101020127-WIRARAB).

(2) Refined **bibliographic integration:** While TEI Lex-0 (and TEI P5) support citation of sources at the dictionary level, this is too coarse-grained for the needs of the SHAWI dictionary. To address this, `<entry>` elements in the SHAWI customization may include a `<listBibl>` element which contains placeholders for records from the VICAV bibliography. This allows for the addition of context-specific bibliographic details (like page numbers or comments) while at the same time avoiding multiplication of bibliographic information.

(3) Extended encoding of **features specific to Arabic varieties:** So far, the TEI Lex-0 specification offers no dedicated mechanism for representing morphological structures characteristic of Semitic languages. The SHAWI customization therefore introduces new attribute values for `@type` on `<gram>` to capture phenomena such as root-based derivation, morphological patterns, and verbal stem classes.

By applying the TEI Lex-0 Schema to dialectological context, the SHAWI Dictionary demonstrates the adaptability of community standards to non-Indo-European linguistic data. It contributes both to the ongoing consolidation of digital lexicographic practices and to the sustainable documentation of previously underdescribed Arabic varieties, giving an example of how TEI-based infrastructures can bridge linguistic research, digital humanities, and language technology.

Keywords: TEI Lex-0; lexicographic modelling; dialect dictionaries; Arabic dialectology

References

Budin, G., Majewski, S. & Moerth, K. (2012). Creating Lexical Resources in TEI P5. *Journal of the Text Encoding Initiative (jTEI)* 3. Available at: https://doi.org/10.4000/jtei.522.

Moerth, K., Schopper, D. & Siam, O. (2015). Towards a Diatopic Dictionary of Spoken Arabic Varieties: Challenges in Compiling the VICAV Dictionaries. In G. Grigore & G. Biţună (eds) *Arabic Varieties: Far and Wide. Proceedings of the 11th International Conference of AIDA*. Bucharest, pp. 395–404.

Procházka, S. & Moerth, K. (2015). The Vienna Corpus of Arabic Varieties: building a digital research environment for Arabic dialects. In M. Al-Hamad, R. Ahmed and H. Aloui (eds.) *Lisan Al-Arab: Studies in Contemporary Arabic Dialects, Proceedings of the 10th International Conference of AIDA*. Qatar University 2013. Vienna: LIT Verlag, pp. 209–218.

Rahtz, S. (2014). Advanced topics in ODD. *TEI Conference Workshop: An Introduction to TEI's ODD: One Document Does it all.* Oct 2014, Evanston, United States. 2014. Available at: https://inria.hal.science/hal-01767683.

Tasovac, T., Romary, L., Banski, P., Bowers, J., Does, J. de, Depuydt, K., Erjavec, T., Geyken, A., Herold, A., Hildenbrandt, V., Khemakhem, M., Lehečka, B., Petrović, S., Salgado, A. & Witt, A. (2018ff.). TEI Lex-0: A baseline encoding for lexicographic data. DARIAH Working Group on Lexical Resources. Available at: https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html.

TEI Consortium. (2025). TEI P5: Guidelines for Electronic Text Encoding and Interchange, P5 Version 4.9.0. Available at: www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf.

Enhancing Lexicographic Access for Deaf and Hard-of-Hearing Learners: A Digital Greek Sign Language Dictionary with AI-Powered Language Support

Isidora Despotidou, Zoe Gavriilidou

Democritus University of Thrace, Greece
E-mail: isidora.despotidou@gmail.com, zoegabriil@gmail.com

Abstract

The purpose of the presentation is to explore the design and development of an innovative online pedagogical dictionary of Greek Sign Language, specifically tailored to the linguistic and educational needs of Deaf and Hard-of-Hearing (DHH) learners in Greece. Emphasizing accessibility and pedagogical usability, the dictionary integrates Artificial Intelligence (AI) technologies to support multimodal interaction and facilitate bilingual proficiency in both Greek and Greek Sign Language (GSL).

Implemented as a web-based platform, the dictionary ensures broad accessibility for secondary and tertiary-level students through an intuitive, learner-centered interface. Key features include:

- An interactive chatbot, enabling users to ask questions via either spoken/written Greek or sign language, receiving responses in both modalities.
- AI-assisted exercise generation, which adapts vocabulary and grammar tasks based on individual learner profiles and performance metrics.
- Neural-network-based text-to-sign translation modules, allowing for real-time rendering of written Greek input into Greek Sign Language.

The presentation is structured in three main parts: it begins with a discussion on the significance of inclusive lexicography and the imperative to develop language resources that address the accessibility needs of diverse user groups. It then outlines the lexicographic protocol adopted for compiling the dictionary, followed by an in-depth description of the platform's functionalities. The final section analyzes the integration of AI technologies and their role in enhancing both linguistic accessibility and pedagogical personalization.

The contribution of the paper is twofold: first, it provides a concrete model of inclusive digital lexicography for sign languages; second, it highlights how AI can be leveraged not merely as a technical enhancement, but as a transformative tool in promoting equitable access to language resources for underrepresented communities.

Keywords: AI, Greek Sign Language; Inclusive Lexicography; Pedagogical Digital Dictionary; Sign Language Lexicography; Accessibility

References

Cohn, T., & Bender, E. M. (2021). AI and the future of language documentation: Perspectives from the field. *Language Resources and Evaluation*, 55(3), 767–786. Available at: https://doi.org/10.1007/s10579-021-09526-9.

Gouws, R. & D. Prinsloo, (2005). *Principles and practice of South African Lexicography.* Sun Press.

Klosa, A. (2013). The lexicographical process (with special focus on online dictionaries). In Hausmann, F.J., Reichmann, O., Wiegand, H.E., & Zgusta, L. (eds.), Wörterbücher. *Ein internationales Handbuch zur Lexikographie.* Supplement Volume: Recent developments with focus on electronic and computational lexicography. Berlin/Boston: De Gruyter, pp. 517–524.

Lew, R. (2011). Online dictionaries of English. In P.A. Fuertes-Olivera & H. Bergenholtz (eds.) *e-Lexicography: The Internet, Digital Initiatives and Lexicography.* London/New York: Continuum, pp. 230–250.

McKee, R., & Vale, M. (2017). Sign Language Lexicography. In P. Hanks & G.-M. de Schryver (eds.), *International Handbook of Modern Lexis and Lexicography.* Springer.

Miller, C. (2006). Sign Language: Transcription, Notation, and Writing. In K. Brown (ed.), *Encyclopedia of Language and Linguistics* (2nd ed.), pp. 328–338. Amsterdam: Elsevier.

Tarp, S. (1). Reflections on Lexicographical User Research. *Lexikos*, 19. Available at: https://doi.org/10.5788/19-0-440.

Zhou, L., & Li, X. (2022). *Artificial Intelligence for Language Learning: A New Frontier in Educational Technology.* Springer.

POSTERS AND DEMOS

A Bazaar Among Cathedrals – Leveraging Wikidata as an Open Marketplace for Lexicographic Data

Gregor Middell

Berlin-Brandenburg Academy of Sciences and Humanities
E-mail: gregor.middell@bbaw.de

Abstract

Eric Raymond's influential essay (Raymond 1999) about the community-based software development as practiced in the Open Source movement vs. the previously dominant, closed, top-down approach mostly preferred in the commercial realm proved also instructive for the Wikiverse. Its flagship project Wikipedia with a comparable approach to knowledge production and dissemination disrupted the market of encyclopedic offerings to the extent that it became the primary source of information in that context, driving previous commercial market leaders out of business. While Wiktionary, the lexicographic equivalent of Wikipedia, did not have the same effect on its established competitors, it has drawn considerable academic interest as a lexical resource, from favorable comparisons to controlled or closed-source resources (Meyer and Gurevych 2010; 2012) over integrations with such resources (McCrae, Montiel-Ponsoda, and Cimiano 2012) to its conversion and augmentation as a comprehensive, multilingual Linked Open Data resource in its own right (Sérasset 2015). The Wikiverse picked up this research-driven development of structured, machine-readable lexical datasets by incorporating lexicographic information in Wikidata (Lindemann 2025), basing the data model in turn on Ontolex Lemon, the lexicon model for ontologies which originated in a research collaboration.

The Digitales Wörterbuch der deutschen Sprache (DWDS) wanted to further explore this relationship between the academic realm on the one hand, with its lexicographic projects more akin to Raymond's cathedrals, and the bazaar-like, dynamic and community-driven approach on the other, which informs the construction of Wikidata's knowledge graph. In January 2023 the DWDS conducted a data donation of about 185,000 German lexemes to Wikidata. In line with previous studies (Kosem et al. 2021), the facts donated to Wikidata comprised lexical information most likely to be liberally licensed by projects like the DWDS (lexical category, written representations, grammatical features), while other copyrighted information (sense glosses, etymology etc.) was deliberately excluded. The poster presents the challenges of this data donation, for example impedances in mapping the different data models, organizing support in the community or overcoming technical obstacles. It also reports on the first results: Since the initial data import two years ago, the German lexeme inventory of Wikidata grew to over 200,000 entries. By now it registers over 550,000 links of those entries to

external lexical resources beside the DWDS, and last but not least over 11,000 community-contributed links to concepts on the sense level, that in turn link to about 175,000 lexemes in other languages.

Keywords: Crowdsourcing; Wikidata; Linked Open Data; OntoLex Lemon

References

Kosem, I., Nimb, S., Tiberius, C., Boelhouwer, B. & Krek, S. (2021). *License to Use: ELEXIS Survey on Licensing Lexicographic Data and Software*. Available at: https://euralex.org/publications/license-to-use-elexis-survey-on-licensing-lexicographic-data-and-software/.

Lindemann, D. (2025). 'Lexikografie Auf Wiki-Plattformen'. *Lexicographica*. (in-press).

McCrae, J., Montiel-Ponsoda, E. & Cimiano, P. (2012). 'Integrating WordNet and Wiktionary with Lemon'. In *Linked Data in Linguistics*, edited by Christian Chiarcos, Sebastian Nordhoff, and Sebastian Hellmann, pp. 25–34. Berlin, Heidelberg: Springer Berlin Heidelberg. Available at: https://doi.org/10.1007/978-3-642-28249-2_3.

Meyer, C.M. & Gurevych, I. (2010). 'Worth Its Weight in Gold or Yet Another Resource — A Comparative Study of Wiktionary, OpenThesaurus and GermaNet'. In *Computational Linguistics and Intelligent Text Processing*, edited by Alexander Gelbukh, pp. 38–49. Berlin, Heidelberg: Springer. Available at: https://doi.org/10.1007/978-3-642-12116-6_4.

———. (2012). 'Wiktionary: A New Rival for Expert-Built Lexicons? Exploring the Possibilities of Collaborative Lexicography'. In *Electronic Lexicography*, edited by Sylviane Granger & Magali Paquot, 0. Oxford University Press. Available at: https://doi.org/10.1093/acprof:oso/9780199654864.003.0013.

Raymond, E. (1999). 'The Cathedral and the Bazaar'. *Knowledge, Technology & Policy* 12(3), pp. 23–49. Available at: https://doi.org/10.1007/s12130-999-1026-0.

Sérasset, G. (2015). 'DBnary: Wiktionary as a Lemon-Based Multilingual Lexical Resource in RDF'. *Semantic Web* 6(4), pp. 355–61. Available at: https://doi.org/10.3233/SW-140147.

Implementing Frames in the *Phrase-based Active Dictionary*: why Frames are needed but FrameNet can only be a partial solution

Laura Rebosio

University of Innsbruck, Department of Translation Studies
Herzog-Siegmund-Ufer 15, 6020 Innsbruck
E-mail: Laura.Rebosio@uibk.ac.at

Abstract

This paper explores the differences between the *Phrase-based Active Dictionary* (PAD) and FrameNet in their approaches to meaning representation, focusing on the verbs *agree* and *follow*. The PAD, a component of the PhraseBase project, adopts a splitting-friendly methodology that emphasizes granularity and ontological consistency, ensuring a more comprehensive coverage of polysemy. In contrast, FrameNet prioritizes broader conceptualization, often leaving finer distinctions unaddressed. Through a detailed matching process, this analysis reveals that several senses traced in the PAD are not covered or not distinguished in FrameNet, highlighting the need for an extended concept of Frame. The proposed extension of the system includes increased granularity, the incorporation of encyclopedic knowledge by using ostensive aids, and cultural sensitivity. These enhancements would improve the visual representation of Frames or enhance their representation potential, making them more accessible and informative for users of the PAD. The paper concludes by addressing open questions about the systematic implementation of these extensions and their implications for linguistic analysis and lexicographic practice. By combining theoretical insights with practical applications, the PAD aims to offer a model for deepening meaning representation for advanced language learners and translators.

Keywords: Phrase-based Active Dictionary (PAD); FrameNet; Frame's application;

encyclopaedic information in lexicography

References

Brugman, C. & Lakoff, G. (1988). Cognitive Topology and Lexical Networks. In S. Small, G. Cottrell & M. Tanenhaus (eds.) *Lexical Ambiguity Resolution: Perspectives from Psycholinguistics, Neuropsychology, and Artificial Intelligence.* San Mateo (CA): Morgan Kaufmann Publishers, pp. 477–508. Available at: https://doi.org/10.1016/B978-0-08-051013-2.50022-7.

DiMuccio-Failla, P. & Giacomini, L. (2017a). Designing a Learner's Dictionary with Phraseological Disambiguators. In R. Mitkov (ed.) *Computational and Corpus-Based Phraseology: Proceedings of the Second International Conference EUROPHRAS 2017*. Cham: Springer, pp. 290–205. Available at: https://doi.org/10.1007/978-3-319-69805-2.

DiMuccio-Failla, P. & Giacomini, L. (2017b). Designing a Learner's Dictionary based on Sinclair's Lexical Units by means of Corpus Pattern Analysis and the Sketch Engine. In I. Kosem et al. (eds.) *Proceedings of the ELEX 2017 Conference.* Brno: Lexical Computing CZ s.r.o, pp. 437–457. Available at: https://elex.link/elex2017/proceedings/eLex_2017_Proceedings.pdf.

DiMuccio-Failla, P. & Giacomini, L. (2022). A Proposed Microstructure for a New Kind of Active Learner's Dictionary. *Lexicographica*, 38(1), pp. 475–499. Available at: https://doi.org/10.1515/lex-2022-0016.

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*. Berlin/Boston: De Gruyter, pp. 19–90. Available at: https://doi.org/10.1515/9783111545943.

Fillmore, C. (1975). An Alternative to Checklist Theories of Meaning. *Proceedings of the First Annual Meeting of the Berkeley Linguistics Society*, pp. 123–131. Available at: https://doi.org/10.3765/bls.v1i0.2315.

FrameNet, Berkeley University. Accessed at: https://framenet.icsi.berkeley.edu/. (20 June 2025)

Giacomini, L. & DiMuccio-Failla, P. (2019). Investigating Semi-Automatic Procedures in Pattern-Based Lexicography. In I. Kosem et al. (eds.) *Proceedings of the ELEX 2019 Conference*. Brno: Lexical Computing CZ s.r.o, pp. 490–505. Available at: https://elex.link/elex2019/wp-content/uploads/2019/09/eLex_2019_28.pdf.

Giacomini, L. (2025). Introduction to the PhraseBase Project. In L. Giacomini & V. Piunno (eds.) *Patterns of Meaning in Lexicography and Lexicology*. Berlin/Boston: De Gruyter, pp. 15–17. Available at: https://doi.org/10.1515/ 9783111545943.

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

Kilgarriff, A. et al. (2004). The Sketch Engine. In G. Williams & S. Vessier (eds.) *Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004. Lorient: Faculté des Lettres et des Sciences Humaines, Université de Bretagne Sud*, pp. 105–116. Available at: https://shorturl.at/b0isb.

Ruppenhofer, J. et al. (2016). *FrameNet II: Extended Theory and Practice*. Available at: https://web.archive.org/web/20221026121837/https://framenet2.icsi.berkeley.edu/docs/r1.7/book.pdf.

Sinclair, J.M. (2004). *Trust the Text: Language, Corpus and Discourse*. London/New York: Routledge.

Svensén, B. (2009). *A Handbook of Lexicography – The Theory and Practice of Dictionary-Making*. Cambridge: Cambridge University Press.

WordNet, Princeton University. Accessed at: https://wordnet.princeton.edu/. (5 July 2025)

Project of a Specialized Dictionary Website

Mykyta Yablochkov, Alona Dorozhynska, Iryna Ostapova,

Iuliia Verbynenko

Ukrainian Lingua-Information Foundation of National Academy of Sciences of Ukraine,
3 Holosiivska avenue, 03039, Kyiv, Ukraine
E-mail: irinaostapova@gmai.com, eugeniokuprianov@gmail.com, gezartos@gmail.com

Abstract

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), 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. 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. 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 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).

Keywords: computer lexicography; lexicographic system; specilised dictionary; parsing; lexicographic database

The Hare and the Tortoise: Pipeline for Latvian Information and Communication Technologies Secondary Term Formation

Dace Šostaka, Inguna Skadiņa

Department of Computing, Faculty of Science and Technology, University of Latvia,
Raiņa bulvāris 19, Riga, Latvia
E-mail: inguna.skadina@lu.lv, dace.sostaka@lu.lv

Abstract

The Information and Communication Technologies (ICT) field has evolved rapidly in recent decades. Thus, to describe new devices, activities, and concepts that appear yearly, a vast number of terms are created primarily in English, while other languages rely on secondary term formation (STF) for ICT end-users (ETSI Guide, 2022). Systematic secondary rendering and dissemination up-to-date terminology in the target language (Chiocchetti and Ralli, 2013; Stefaniak, 2023) are crucial for language development and benefit professionals, students, and the public. We analysed the STF process in Latvian for the ICT domain during the development of the Language Technology (LT) course at the University of Latvia.

For over 30 years, the Terminology Commission of the Latvian Academy of Sciences (TCLAS, 2025) and its sub-commissions, including the Information and Communication Technologies Sub-Commission (ICTSC), have carried out term formation. ICTSC comprises of ICT professionals, terminologists, and linguists. ICT students also participate in meetings to approbate terms for the first time. The commission meets twice a month during the academic year. Terms are sourced from higher education, industry, and translating agencies, including the European Commission. They are added to the biweekly agenda, discussed, and, if accepted, recorded in an open-access Academic term database, available on the web since 2005 (ATB, 2025).

For the LT course, terms were manually extracted from lecture slides. Given ICTSC's capacity to produce about 20 high-quality terms during a 2-hour meeting, terms were prioritised based on their relevance in the LT course. Identified terms were reviewed and defined, supplemented with usage examples and visuals. Possible Latvian term variants were proposed, with ICTSC members conducting preliminary written discussions, and 111 terms were accepted and are available in the Academic term database (ATB, 2025).

The STF process includes several challenges where AI tools could be applied. As the concept of the term is usually expressed most precisely in its definition, the most significant challenge is providing a clear definition for terms used in several ICT

subdomains. Second comes weighing arguments for and against creating source-language oriented terms that can be easily back-translated and will be recognisable versus creating secondary terms that precisely reflect the definition but might be far from the direct translation of the original term (e.g., Bag of Words). The third challenge is the length of the term and euphonism – how easily it can be pronounced. As a rule of thumb, the longer the term, the less likely it will be used in spoken communication, and the direct calque will be used.

The STF process was researched (Šostaka et al., 2023), and several approaches were tested to speed up "mechanical" parts of the term creation. The first approach was using an AI tool (ChatGPT 4.0) on 140 concepts and terminology units within *ISO/IEC 22989:2022(en)*, searching and then evaluating suggestions for STF in Latvian and comparing them to the terms already approved by the Terminology Commission (Šostaka et al., 2025). Out of 140 concepts, 75 terms had an exact match, 65 had a partial match, while 5 had no match.

The second approach was checking the time saved using a tool for term extraction from online dictionaries (Šostaka et al., 2024). The tool allows to review user-specified sources (e.g., Merriam-Webster dictionary) on the Internet, related to ICT terms; it is scalable, and it is possible to add sources of the user's choice in other fields and languages. It allowed us to save 74 minutes when searching 40 terms, as opposed to 106 minutes needed for a manual search.

Keywords: Secondary term formation process; Information and Communication Technologies; Language Technology; Artificial Intelligence Tools; Latvian

References

ATB, Akadēmiskā terminu datubāze [Academic term database]. (2025). Accessed at: http://www.akadterm.lv. (19 March 2025)

Chiocchetti Elena & Ralli Natascia (editors). (2013). *Guidelines for collaborative legal/administrative terminology work*, pp. 36–39. Available at: https://cordis.europa.eu/docs/projects/cnect/7/270917/080/deliverables/001-D33Guidelinesforcollaborativelegaladministrativeterminologywork.pdf. (18 March 2025)

ETSI Guide EG 203 499. (2022). *ETSI simplifies ICT end-users' lives with a guide available in 19 European languages*. Accessed at: https://www.etsi.org/newsroom/press-releases/2105-2022-07-etsi-simplifies-ict-end-users-lives-with-a-guide-available-in-19-european-languages. (19 March 2025)

Stefaniak, K. (2023). *Terminology management and terminology quality assurance in the European Commission's Directorate-General for Translation*, Handbook of Terminology, pp. 351–374. Available at: https://doi.org/10.1075/ hot.3.ter12. (18 March 2025)

Šostaka, D., Borzovs, J., Cauna, E., Keiša, I., Pinnis, M. & Vasiļjevs, A. (2023). The Semi-Algorithmic Approach to Formation of Latvian Information and

Communication Technology Terms. *Baltic Journal of Modern Computing*, 11(1), pp. 67–89. Available at: https://doi.org/10.22364/bjmc 2023.11.1.05.

Šostaka, D., Borzovs, J. & Zuters, J. (2023). *Preliminary Results on the Use of Artificial Intelligence in Secondary Term Formation*. 14th International Conference on Data Analysis Methods for Software Systems. Druskininkai, Lithuania. Available at: https://drive.google.com/file/d/12a3lZpvv7SxzBo2B03zu1fhdePinCy9n/view. (19 March 2025)

Šostaka, D., Sondors, K., Borzovs, J. & Zuters J. (2024). *Information Retrieval Tool for Secondary Term Creating: Preliminary Results*. 15th International Conference on Data Analysis Methods for Software Systems. Druskininkai, Lithuania. Available at: https://www.mii.lt/damss/files/posters_2024/id_I-9.pdf. (19 March 2025)

TCLAS, Latvijas Zinātņu akadēmijas Terminoloģijas komisija [Terminology Commission of the Latvian Academy of Sciences]. Accessed at: https://www.lza.lv/par-mums/terminologijas-komisija. (18 March 2025)

Exploring Derivational Families through Intelligent Lexicography

Krešimir Šojat[1], Kristina Kocijan[2]

[1] Department of Linguistics, Faculty of Humanities and Social Sciences, University of Zagreb, Ivana Lučića 3 Zagreb Croatia

[2] Department of Information and Communication Sciences, Faculty of Humanities and Social Sciences, University of Zagreb, Ivana Lučića 3 Zagreb Croatia

E-mail: ksojat@ffzg.unizg.hr, krkocijan@ffzg.unizg.hr

Abstract

This paper presents a novel approach to exploring derivational families within the framework of Intelligent Lexicography, using the ŠKOLARAC corpus: a collection of Croatian school essays written by L1 learners (native-speaking students) in grades 5 through 8 and enriched with metadata such as gender, grade level, and region. By combining rule-based linguistic processing in NooJ, a linguistic development environment for formalizing morphological and syntactic patterns, with tailored morphological procedures for Croatian, the study identifies and maps derivational networks of three pedagogically relevant lexical morphemes (CRT, PIS, and RAD) tracing their associated inflected and derived forms as they appear in young learner corpora. The extracted data are visualized using radial graphs, butterfly charts, and hierarchical structures, enabling a multifaceted analysis of morphological productivity and lexical variation. This integrated workflow demonstrates how intelligent tools can enhance lexicographic practice by uncovering deep morphological relationships in authentic learner language. The findings support the development of adaptive, learner-sensitive lexicographic resources with applications in linguistics, language education, and curriculum design, particularly in the context of developing digital dictionaries and vocabulary tools tailored to young learners.

Keywords: intelligent lexicography; derivational families; learner corpora; Croatian morphology; linguistic visualization; ŠKOLARAC corpus

References

Babić, S. (2002). *Tvorba riječi u hrvatskome književnome jeziku.* Zagreb: HAZU, Globus.

Barić, E., Lončarić, M., Malić, D., Pavešić, S., Peti, M., Zečević, V. & Znika, M. (2003). *Hrvatska gramatika.* Zagreb: Školska knjiga.

Filko, M., Šojat, K. & Štefanec, V. (2020). The design of CroDeriv 2.0. *The Prague*

Bulletin of Mathematical Linguistics, 115, pp. 83–104.

Frankenberg-Garcia, A. (2014). How language learners can benefit from corpora, or not. In *Recherches en didactique des langues et des cultures [En ligne]*, 11(1). Available at: https://doi.org/10.4000/rdlc.1702.

Gabrielatos, C. (2005). "Corpora and language teaching: just a fling or wedding bells?". *Teaching English as a Second Language Electronic Journal*, 8(4). pp. 1–35. Available at: http://tesl-ej.org/ej32/a1.html.

Hržica, G. (2021). Derivational morphology in Croatian child language. In *The Acquisition of Derivational Morphology. A Cross-linguistic Perspective*. Amsterdam: John Benjamins Publishing, pp. 141–168. Available at: https://doi.org/10.1075/lald.

Kocijan, K. (2015). Visualizing natural language resources. In F. Pehar, C. Schlögl & C. Wolff (eds.) *Proceedings of the 14th International Symposium on Information Science (ISI 2015)*. Zagreb: Verlag Werner Hülsbusch, pp. 203–216. Available at: https://zenodo.org/record/17934/files/s3_203-216.pdf.

Kocijan, K. (2022). How we color the world with words. *Suvremena lingvistika*, 48(93), pp. 41–83. Available at: https://doi.org/10.22210/suvlin.2022.093.03.

Kocijan, K., Janjić, M. & Librenjak, S. (2016). Recognizing diminutive and augmentative Croatian nouns. In *Automatic Processing of Natural-Language Electronic Texts with NooJ*. Cham: Springer, pp. 23–36.

Kocijan, K. & Šojat, K. (2024). Exposing diminutive and pejorative verbs in Croatian. In A. Bartulović, L. Mijić & M. Silberztein (eds.) *Formalizing Natural Languages: Applications to Natural Language Processing and Digital Humanities*. Cham: Springer, pp. 39–51. Available at: https://doi.org/10.1007/978-3-031-89810-5_4.

Kocijan, K., Šojat, K. & Poljak, D. (2018). Designing a Croatian aspectual derivatives dictionary: preliminary stages. In *Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing (LR4NLP-2018)*. Stroudsburg (PA): Association for Computational Linguistics (ACL), pp. 28–37.

Kuna, Z. (2022). Tvorba imenica u ranom jezičnom razvoju na temelju podataka za jedno dijete iz Hrvatskog korpusa dječjeg jezika. *Suvremena lingvistika*, 48. Available at: https://doi.org/10.22210/suvlin.2022.093.04.

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

Llach, M.P.A. (2010). Exploring the role of gender in lexical creations. In R.M.J. Catalán (ed.) *Gender Perspectives on Vocabulary in Foreign and Second Languages*. London: Palgrave Macmillan. Available at: https://doi.org/10.1057/9780230274938_4.

Murakami, A. & Alexopoulou, T. (2016). L1 influence on the acquisition order of English grammatical morphemes: A learner corpus study. *Studies in Second Language Acquisition*, 38(3), pp. 365–401. Available at: https://doi.org/10.1017/S0272263115000352.

Quixal, M., Rudzewitz, B., Bear, E. & Meurers, D. (2021). Automatic annotation of

curricular language targets to enrich activity models and support both pedagogy and adaptive systems. In *Proceedings of the 10th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2021)*. Linköping Electronic Conference Proceedings, 177, pp. 15–27.

Römer-Barron, U. (2023). Usage-based approaches to second language acquisition vis-à-vis data-driven learning. *TESOL Quarterly*, 58. Available at: https://doi.org/10.1002/tesq.3278.

Silić, J. & Pranjković, I. (2005). *Gramatika hrvatskoga jezika: za gimnazije i visoka učilišta*. Zagreb: Školska knjiga.

Šojat, K., Kocijan, K. & Filko, M. (2019). Processing Croatian aspectual derivatives. In M. Mirto et al. (eds.) *Formalizing Natural Languages with NooJ 2018 and Its Natural Language Processing Applications*. Heidelberg: Springer, pp. 50–61.

Šojat, K. & Filko, M. (2023). Processing Croatian morphology: Roots, segmentation and derivational families. In M. Filko & K. Šojat (eds.) *Proceedings of the 4th International Workshop on Resources and Tools for Derivational Morphology*. Zagreb: HDJT, pp. 61–70.

Šojat, K., Srebačić, M. & Tadić, M. (2012). Derivational and semantic relations of Croatian verbs. *Journal of Language Modelling*, 0(1), pp. 111–142.

Šojat, K., Srebačić, M., Tadić, M. & Pavelić, T. (2014). CroDeriV: A new resource for processing Croatian morphology. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik: ELRA, pp. 3366–3370.

Tadić, M. & Fulgosi, S. (2003). Building the Croatian morphological lexicon. In *Proceedings of the EACL2003 Workshop on Morphological Processing of Slavic Languages*. Budapest: ACL, pp. 41–46.

Teibowei, M.T. (2024). Sociolinguistic variations and gender differences in language usage. *International Journal of English Language and Communication Studies*, 9(1), pp. 95–101.

Vučković, K., Librenjak, S. & Dovedan, Z. (2011). Deriving nouns from numerals. In *Proceedings of the NooJ 2010 International Conference and Workshop*. Komotini, pp. 84–95.

Vučković, K., Librenjak, S. & Dovedan Han, Z. (2013). Derivation of adjectives from proper names. In *Formalising Natural Languages with NooJ*. Newcastle upon Tyne: Cambridge Scholars Publishing, pp. 57–68.

Vučković, K., Tadić, M. & Bekavac, B. (2010). Croatian language resources for NooJ. *CIT. Journal of Computing and Information Technology*, 18, pp. 295–301. Available at: https://doi.org/10.2498/cit.1001914.

Comparative Analysis of Medical Adjectives in Croatian General Dictionaries

Martina Pavić[1], Daša Farkaš[2]

[1] Institute of Croatian Language, Republike Austrije 16, 10000 Zagreb
[2] Faculty of Humanities and Social Sciences, Ivana Lučića 3, 10000 Zagreb
E-mail: mpavic@ihjj.hr, dfarkas@ffzg.hr

Abstract

The representation of medical adjectives in Croatian general dictionaries reveals significant inconsistencies, reflected in uneven lemma inclusion, ambigous or absent domain labels, and limited definitional precision. This paper analyzes the 80 most frequent adjectives, based on corpus data from the *Croatian Medical Corpus* (CMC) (Kocijan, Kurolt & Mijić, 2020), in the three major Croatian general dictionaries: *Veliki rječnik hrvatskoga standardnog jezika* (2015), *Hrvatski enciklopedijski rječnik* (2002), and *Rječnik hrvatskoga jezika* (2000). The analysis focuses on lemma status, the presence of domain labels, and the accuracy of definitions.

To contextualize the Croatian practice, the study includes a brief comparison with *Merriam-Webster Dictionary* (2025), which demonstrates better lemma coverage and more terminologically informed definitions, but also exhibits inconsistencies that reflect the broader challenges of systematically representing medical adjectives in general lexicography.

The paper's findings reveal inconsistencies in Croatian lexicographic practice and highlight the need for more conceptually grounded, corpus-based approaches that integrate terminological precision with lexicographic usability.

Keywords: medical adjectives; Croatian general dictionaries; *Croatian Medical Corpus*; *Merriam-Webster Dictionary*

References

Books:

Alberts, M. (2001). Lexicography versus Terminograpy. *Lexikos*, 11, pp. 71–84.

Alonso Campos, A. & Torner Castells, S. (2010). Adjectives and collocations in specialized texts: lexicographic implications, In A. Dykstra & T. Schoonheim (eds.) *Proceedings of the XIV EURALEX International Congress*, Leeuwarden: Fyrske Akademy – Afûk, pp. 872–881.

Bergenholtz, H. & Tarp, S. (eds.) (1995). *Manual of Specialised Lexicography: The preparation of specialized dictionaries*. Amsterdam/Philadelphia: John Benjamins.

Cabré, M.T. (1999). *Terminology: Theory, methods, and applications*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Costa, R. (2013). Terminology and Specialised Lexicography: two complementary domains. *Lexicographica*, 29(1), pp. 29–42.

Cabré Castellvi, M.T. (2003). Theories of terminology: Their description, prescription and explanation. *Terminology*, 9(2), pp. 163–199.

Durán-Muñoz, I. (2019). Adjectives and their keyness: a corpus-based analysis of tourism discourse in English. *Corpora*, 14(3), pp. 351–378.

Fontenelle, T. (2016). From Lexicography to Terminology: a Cline, not a Dichotomy. In A. Abel, C. Vettori & N. Ralli (eds.) *Proceedings of the XVI EURALEX International Congress: The User in Focus*, Bolzano: Institute for Specialised Communication and Multilingualism, pp. 25–45.

Grčić Simeunović, L. (2015). Prilog metodologiji opisa sintagmi u stručnom diskursu. *Rasprave: Časopis Instituta za hrvatski jezik i jezikoslovlje*, 41(1), pp. 29–47.

Grčić Simeunović, L., Stepišnik, U. & Vintar, Š. (2020). Klasifikacijska uloga pridjeva u domeni geomorfologije krša. *Rasprave: Časopis Instituta za hrvatski jezik i jezikoslovlje*, 46(2), pp. 619–633.

Grčić, Simeunović, L. (2021). *Terminološki opis u službi stručnoga prevođenja. Dinamično modeliranje specijaliziranoga znanja*. Zadar/Zagreb: Sveučilište u Zadru/Institut za hrvatski jezik i jezikoslovlje.

ISO 704:2009. *Terminology work – Principles and methods*. Geneva: International Organization for Standardization.

ISO 1087:2019. *Terminology Work – Vocabulary – Part 1: Theory and Application*. Geneva: International Organization for Standardization.

Jojić, L. & Matasović, R. (2000). *Hrvatski enciklopedijski rječnik* [Croatian Encyclopedic Dictionary]. Zagreb: Novi liber.

Jojić, L. (2015). *Veliki rječnik hrvatskoga standardnog jezika* [Great Dictionary of the Croatian Standard]. Zagreb: Školska knjiga.

Kocijan, K., Kurolt, S. & Mijić, L. (2020). Building the Croatian medical dictionary from medical corpus. *Rasprave: Časopis Instituta za hrvatski jezik i jezikoslovlje*, 46(2), pp. 765–782.

Kilgarriff, A. et al. (2014). *The Sketch Engine: Ten years on*. Lexicography, 1(1), pp. 7–36.

Landau, S.I. (2001). *Dictionaries. The art and craft of lexicography*. Cambridge: Cambridge University Press.

Mel'čuk, I. & Polguère, A. (2018). Theory and practice of lexicographic definition. *Journal of Cognitive Science,* 19(4), pp. 417–470.

Merriam-Webster. Accessed at: https://www.merriam-webster.com. (1–8 July 2025)

Pitkänen-Heikkilä, K. (2015). Adjectives as terms. *Terminology*, 21(1), pp. 76–101.

Sager, J.C. (1990). *A practical course in terminology processing*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Salgado, A. (2021). *Terminological Methods in Lexicography: Conceptualising, Organising and Encoding Terms in General Language Dictionaries* (Doctoral dissertation). Universidade Nova de Lisboa, Faculdade de Ciências Sociais e Humanas, Lisboa.

Salgado, A., Costa, R. & Tasovac, T. (2022). Applying terminological methods to lexicographic work: terms and their domains. In A. Klosa-Kückelhaus et al. (eds.) *Dictionaries and Society. Proceedings of the XX EURALEX International Congress*, Mannheim: IDS-Verlag, pp. 181–195.

Salgado, A. & Costa, R. (2024). Enhancing Lexicographic Work with Terminological Methods. In T. Margalitadze (ed.) *Proceedings of the I International Conference Lexicography in the XXI century: Lexicography – by combining traditional methods and modern technologies*, Tbilisi: Centre for Lexicography and Language Technologies, Ilia State University, pp. 15–26.

Šonje, J. (2000). *Rječnik hrvatskoga jezika* [Dictionary of the Croatian Language]. Zagreb: Leksikografski zavod Miroslav Krleža.

Temmerman, R. (2000). *Towards New Ways of Terminology Description. The Sociocognitive Approach.* Amsterdam/Philadelphia: John Benjamins Publishing Company.

Wüster, E. (1979). *Einführung in Die Allgemeine Terminologielehre Und Terminologische Lexikographie*. Beč/New York: Springer International Publishing.

CJVT Igre: New Word Games Based on the Digital Dictionary Database of Slovene

Špela Arhar Holdt[1,2], Iztok Kosem[1,2,3]

[1] Faculty of Arts, University of Ljubljana, Aškerčeva 2, 1000, Ljubljana, Slovenia
[2] Faculty of Computer and Information Science, University of Ljubljana, Večna pot 113, 1000, Ljubljana, Slovenia
[3] Institut "Jožef Stefan", Jamova cesta 39, 1000, Ljubljana, Slovenia
E-mail: Spela.ArharHoldt@ff.uni-lj.si, Iztok.Kosem@fri.uni-lj.si

Abstract

CJVT igre (https://igre.cjvt.si/) is a new digital platform offering word games designed to foster lexical awareness and engagement with standard Slovene. Developed by the Centre for Language Resources and Technologies at the University of Ljubljana, the portal currently hosts three games—Cvetka, Besedolov, and Vezalka—with two more in development. Each game utilizes curated lexical data from the Digital Dictionary Database of Slovene, enhanced through targeted lexicographic work to ensure playability, thematic coherence, and age-appropriateness. This includes refining word lists, rating difficulty, and enriching entries with semantic metadata. Cvetka focuses on orthographic guessing tasks with daily thematic prompts, Besedolov on semantic word search challenges within 11x11 grids, and Vezalka on word formation from a constrained letter set. Designed for both educational and general audiences, the games integrate varying levels of difficulty, optional hints, and dynamic scoring. This paper showcases the platform's interface, gameplay mechanics, and the linguistic and technical adaptations required to transform lexicographic resources into effective digital games.

Keywords: language games; Digital Dictionary Database; semantic type; lexicon;

Slovene

References

Arhar Holdt, Š., Logar, N., Pori, E. & Kosem, I. (2021). Game of Words: Play the Game, Clean the Database. In *Proceedings of the 14th Congress of the European Association for Lexicography (EURALEX 2021)*. Alexandroupolis, Greec, pp. 41–49.

Arhar Holdt, Š., Gantar, P., Kosem, I., Pori, E., Robnik Šikonja, M. & Krek, S. (2023). Thesaurus of Modern Slovene 2.0. In *Proceedings of eLex 2023: Electronic Lexicography in the 21st Century*. Brno, pp. 366–381. Available at: https://elex.link/elex2023/wp-content/uploads/82.pdf.

Bustrillo, H.T.B., Sumicad, R. & Cuevas, G.C. (2024). Effectiveness of Word Games in Teaching Students Vocabulary. *Journal of World Englishes and Educational Practices*, 6(3), pp. 07–16. Available at: https://doi.org/10.32996/jweep.2024.6.3.2.

Carr, M. (1997). Internet Dictionaries and Lexicography. *International Journal of Lexicography*, 10(3), pp. 209–230.

Derakhshan, A. & Khatir, E.D. (2015). The effects of using games on English vocabulary learning. *Journal of Applied Linguistics and Language Research*, 2(3), pp. 39–47.

Dharmayasa, I.N.W. (2022). An Implementing the Hangman Game in Teaching English Vocabulary to Elementary School Students. *Jurnal Pendidikan Bahasa Inggris Undiksha*, 10(3), pp. 291–297. Available at: https://doi.org/10.23887/jpbi.v10i3.58140.

Dobrovoljc, K., Krek, S. & Erjavec, T. (2018). The Sloleks Morphological Lexicon and its Future Development. In *Dictionary of Modern Slovene: Problems and Solutions*. Ljubljana: Znanstvena založba Filozofske fakultete, pp. 42–63.

Friðriksdóttir, S.R. & Einarsson, H. (2022). Fictionary-Based Games for Language Resource Creation. In *Proceedings of the 2nd Workshop on Novel Incentives in Data Collection from People: models, implementations, challenges and results within LREC 2022*, pp. 25–31.

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

Gantar, P., Kosem, I. & Krek, S. (2016). Discovering automated lexicography: the case of Slovene lexical database. *International Journal of Lexicography*, 29(2), pp. 200–225. Available at: https://doi.org/10.1093/ijl/ecw014.

Genovese, F., Bolognesi, M.M., Di Iorio, A. & Vitali, F. (2024). The advantages of gamification for collecting linguistic data: A case study using Word Ladders. *Online Journal of Communication and Media Technologies*, 14(2), e202426. Available at: https://doi.org/10.30935/ojcmt/14443.

Guillaume, B., Fort, K. & Lefebvre, N. (2016). Crowdsourcing Complex Language Resources: Playing to Annotate Dependency Syntax. In *Proceedings of the International Conference on Computational Linguistics (COLING)*. Available at: https://hal.inria.fr/hal-01378980/.

Halimah, F. & Izzah, L. (2020). Building FL-Vocabulary Transferability through Semantic Boggle. *English Language in Focus (ELIF)*, 2(2), pp. 79–86.

Hidayat, N. (2016). Improving Students' Vocabulary Achievement through Word Game. *JEES (Journal of English Educators Society)*, 1(2), pp. 95–104. Available at: https://doi.org/10.21070/jees.v1i2.446.

Khusaini, F. & Fauziah, N. (2024). The Implementation of Word Games to Improve Students' English Vocabulary Proficiency. *Journal of Classroom Action Research*, 6(3), pp. 592–595. Available at: https://doi.org/10.29303/jcar.v6i3.8746.

Klemenc, B., Robnik-Šikonja, M., Fürst, L., Bohak, C. & Krek, S. (2017). Technological

Design of a State-of-the-art Digital Dictionary. In V. Gorjanc, P. Gantar, I. Kosem & S. Krek (eds) *Dictionary of modern Slovene: problems and solutions.* Ljubljana: Ljubljana University Press, Faculty of Arts, pp. 10–22.

Kosem, I. & Pori, E. (2021). Slovenske ontologije semantičnih tipov: samostalniki. In I. Kosem (ed.) *Kolokacije v slovenščini*. Ljubljana: Znanstvena založba Filozofske fakultete, pp. 159–202.

Kosem, I., Krek, S. & Gantar, P. (2021). Semantic data should no longer exist in isolation: the Digital Dictionary Database of Slovenian. In Z. Gavriilidou, L. Mitits & S. Kiosses (eds.) *EURALEX XIX, Congress of the European Association for Lexicography, Lexicography for inclusion: 7–9 September 2021, virtual: abstracts book*. Komotini: Democritus University of Thrace, pp. 81–83. Available at: https://euralex2020.gr/wp-content/uploads/2021/09/EURALEX2020_BookOfAbstracts-Preview-1.pdf.

Kosem, I., Arhar Holdt, Š., Gantar, P. & Krek, S. (2023). Collocations Dictionary of Modern Slovene 2.0. In *Proceedings of eLex 2023: Electronic Lexicography in the 21st Century*, Brno, pp. 491–507. Available at: https://elex.link/elex2023/wp-content/uploads/100.pdf.

Kosem, I., Bálint Čeh, J., Ponikvar, P., Zaranšek, P., Kamenšek, U., Koša, P., Grósf, A., Gantar, P., Krek, S., Arhar Holdt, Š. & Gorjanc, V. (eds.) (2024). *Veliki slovensko-madžarski slovar = Szlovén-magyar nagyszótár*. Ljubljana: Založba Univerze. Accessed at: https://doi.org/10.4312/9789612973049.

Zingano Kuhn, T., Arhar Holdt, Š., Kosem, I., Tiberius, C., Koppel, K. & Zviel-Girshin, R. (2022). Data preparation in crowdsourcing for pedagogical purposes: the case of the CrowLL game. *Slovenščina 2.0: Empirične, aplikativne in interdisciplinarne raziskave*, 10(2), pp. 62–100. Available at: https://doi.org/10.4312/slo2.0.2022.2.62-100.

Kwan, Y.H. (2023). Guess the Right Word: A Review of the Language Game Wordle. *RELC Journal*, 55(3), pp. 855–859. Available at: https://doi.org/10.1177/00336882231192121.

Lafourcade, M. (2007). Making people play for Lexical Acquisition with the JeuxDeMots prototype. In *Proceedings of the 7th International Symposium on Natural Language Processing (SNLP'07)*. Pattaya, Thailand.

Lafourcade, M., Joubert, A. & Le Brun, N. (2015). Collecting and evaluating lexical polarity with a game with a purpose. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pp. 329–337.

Mihaljević, J. (2019). Gamification in E-Lexicography. *INFuture*, pp. 155–164.

Mihaljević, J. & Hudeček, L. (2022). Model for developing educational games based on data from dictionary structure. *Studia Lexicographica*, 16(30), pp. 111–133. Available at: https://doi.org/10.33604/sl.16.30.6.

Munikasari, M., Sudarsono, S. & Riyanti, D. (2021). The Effectiveness of Using Hangman Game to Strengthen Young Learners' Vocabulary. *Journal of English Education Program*, 2(1), pp. 57–65. Available at: https://jurnal.untan.ac.id/index.php/JEEP/article/view/57-65/0.

Orawiwatnakul, W. (2017). El Uso de Crucigramas como Herramienta en el Desarrollo del Vocabulario. *Electronic Journal of Research in Educational Psychology*, 11(30), pp. 413–428. Available at: https://doi.org/10.14204/ejrep.30.12186.

Pamungkas, N.A.R. (2021). The Effects of Wordle Media on Students' Vocabulary Mastery. *JETAL: Journal of English Teaching & Applied Linguistic*, 2(2), pp. 56–61. Available at: https://doi.org/10.36655/jetal.v2i2.531.

Poesio, M., Chamberlain, J., Kruschwitz, U., Robaldo, L. & Ducceschi, L. (2013). Phrase Detectives: Utilizing Collective Intelligence for Internet-scale Language Resource Creation. *ACM Transactions on Interactive Intelligent Systems*, 3(3), pp. 1–44.

Schoonheim, T., Tiberius, C., Niestadt, J. & Tempelaars, R. (2012). Dictionary use and language games: Getting to know the dictionary as part of the game. In *Proceedings of the 15th EURALEX International Congress*. Oslo: Department of Linguistics and Scandinavian Studies, University of Oslo, pp. 974–979.

Toma, I., Alexandru, C.E., Dascalu, M., Dessus, P. & Trausan-Matu, S. (2017). Semantic Boggle: A Game for Vocabulary Acquisition. In É. Lavoué, H. Drachsler, K. Verbert, J. Broisin & M. Pérez-Sanagustín (eds.) *Data Driven Approaches in Digital Education. EC-TEL 2017*. Lecture Notes in Computer Science, vol. 10474. Cham: Springer, pp. 603–607. Available at: https://doi.org/10.1007/978-3-319-66610-5_73.

Handling abstract constructions in a dictionary-based constructicon

Bálint Sass[1], Éva Dömötör[1], Balázs Indig[2], Mátyás Lagos Cortes[3], Veronika Lipp[1], Márton Makrai[4], Gergely Pethő[5]

[1] ELTE Research Centre for Linguistics, Institute for Lexicology
[2] ELTE University, Faculty of Informatics
[3] ELTE Research Centre for Linguistics, Institute for General and Hungarian Linguistics
[4] HUN-REN Research Centre for Natural Sciences, Institute of Cognitive Neuroscience and Psychology
[5] University of Debrecen
E-mail: sass.balint@nytud.hu, domotor.eva@nytud.hu, indig.balazs@inf.elte.hu, lagos.matyas@nytud.hu, lipp.veronika@nytud.hu, makrai.hlt@gmail.com, pagstudium@gmail.com

Abstract

Taking seriously the common construction grammar statement that "it's constructions all the way down" (Goldberg, 2006: 18), the Hungarian Constructicon aims to encompass the widest possible range of constructions. As it is a dictionary-based constructicon, it naturally contains what a dictionary can provide — from morphemes to words, and to partially schematic multiword constructions containing open slots. What had been missing were the more schematic abstract constructions. In this paper, we have added some important constructions of this kind to the database of the constructicon as an experiment, and have enhanced the integrated analyzer tool to handle them appropriately. Now, the system has the machinery to recognize all types of constructions in text and display them to the user. Thanks to the integration of abstract constructions, it does not present constructions in isolation; it reveals the intertwined nature of them, their connections and interactions instead. This results in a fundamentally extended functionality compared to a dictionary. A case study in Section 5 demonstrates the capabilities of the system. The list of the integrated abstract constructions is far from complete, expanding it remains future work.

Keywords: construction; constructicon; abstract construction; constructional schema; lexicon-grammar continuum

References

Bast, R., Endresen, A., Janda, L.A., Lund, M., Lyashevskaya, O., Mordashova, D., Nesset, T., Rakhilina, E., Tyers, F.M. & Zhukova, V. (2021). The Russian Constructicon. An electronic database of the Russian grammatical constructions. Available at: https://constructicon.github.io/russian.

Boas, H.C. (2010). The syntax–lexicon continuum in Construction Grammar: A case study of English communication verbs. *Belgian Journal of Linguistics*, 24(1), pp. 54–82.

De Marneffe, M.C., Manning, C.D., Nivre, J. & Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, 47(2), pp. 255–308. Available at: https://aclanthology.org/2021.cl-2.11/.

Diessel, H. (2023). *The Constructicon*. Cambridge University Press.

Fillmore, C.J., Lee-Goldman, R. & Rhomieux, R. (2012). The FrameNet Constructicon. In H.C. Boas & I.A. Sag (eds.) *Sign-Based Construction Grammar*. Stanford: CSLI Publications, pp. 309–372.

Goldberg, A.E. (2006). *Constructions at work: The nature of generalization in language*. Oxford University Press.

Hilpert, M. (2014). *Construction Grammar and Its Application to English*. Edinburgh University Press.

Janda, L., Endresen, A., Zhukova, V., Mordashova, D. & Rakhilina, E. (2020). How to build a constructicon in five years: The Russian example. *Belgian Journal of Linguistics*, 34, pp. 162–175.

Lorenzi, A., Ljunglöf, P., Lyngfelt, B., Timponi Torrent, T., Croft, W., Ziem, A., Böbel, N., Bäckström, L., Uhrig, P. & Matos, E.E. (2024). MoCCA: A Model of Comparative Concepts for Aligning Constructicons. In H. Bunt, N. Ide, K. Lee, V. Petukhova, J. Pustejovsky & L. Romary (eds.) *Proceedings of the 20th Joint ACL - ISO Workshop on Interoperable Semantic Annotation @ LREC-COLING 2024*. Torino, Italia: ELRA and ICCL, pp. 93–98. Available at: https://aclanthology.org/2024.isa-1.12/.

Lyngfelt, B., Borin, L., Ohara, K. & Torrent, T.T. (eds.) (2018a). *Constructicography: Constructicon development across languages*. Amsterdam: John Benjamins.

Lyngfelt, B., Bäckström, L., Borin, L., Ehrlemark, A. & Rydstedt, R. (2018b). Constructicography at work: Theory meets practice in the Swedish Constructicon. In Lyngfelt et al. (2018a) *Constructicography*, pp. 41–106.

Pannitto, L., Bernasconi, B., Busso, L., Pisciotta, F., Rambelli, G. & Masini, F. (2024). Annotating Constructions with UD: the experience of the Italian Constructicon.

Sass, B. (2024). The "Dependency Tree Fragments" Model for Querying a Constructicon. In K.Š. Despot, A. Ostroški Anić & I. Brač (eds.) *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress*. Cavtat: Institut za hrvatski jezik, pp. 275–283.

Straka, M., Hajič, J. & Straková, J. (2016). UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk & S. Piperidis (eds.)

Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). Portorož, Slovenia: European Language Resources Association (ELRA), pp. 4290–4297. Available at: https://aclanthology.org/L16-1680.

Vainik, E., Paulsen, G., Sahkai, H., Kallas, J., Tavast, A. & Koppel, K. (2024). From a Dictionary to a Constructicon: Putting the Basics on the Map. In K.Š. Despot, A. Ostroški Anić & I. Brač (eds.) *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress*. Institute for the Croatian Language, pp. 209–216.

Ziem, A., Flick, J. & Sandkühler, P. (2019). The German Constructicon Project: Framework, methodology, resources. *Lexicographica*, 35(2019), pp. 15–40.

Qualitative Evaluation of LLM Translation of MWEs for Developing a Croatian Sense Repository

Ana Ostroški Anić[1], Jaka Čibej[2], Ivana Filipović Petrović[3],

Martina Pavić[1], Siniša Runjaić[1], Robert Sviben[1]

[1] Institute for the Croatian Language
[2] University of Ljubljana
[3] Linguistic Research Institute, Croatian Academy of Sciences and Arts
E-mail: aostrosk@ihjj.hr, jaka.cibej@ff.uni-lj.si, ivana.filipov@gmail.com, mpavic@ihjj.hr, srunjaic@ihjj.hr, rsviben@ihjj.hr

Abstract

As part of the COST Action CA21167 *Universality, Diversity and Idiosyncrasy in Language Technology* (UniDive), the ELEXIS-WSD Parallel Sense-Annotated Corpus (Martelli et al., 2021; Čibej et al., 2025) is being expanded to include subcorpora in additional languages—among them, Croatian—as well as new annotation layers. Each language subcorpus of ELEXIS-WSD contains the same 2,024 sentences extracted from WikiMatrix (Schwenk et al., 2019).

The corpus was initially translated from English using two machine translation platforms: Google Translate and Hrvojka (https://hrvojka.gov.hr/). The translations then underwent a two-step manual validation process to first select the more suitable translation for each sentence and correct errors, then the final versions were reviewed in terms of the accuracy of term equivalents and idiomatic expressions. The resulting set was then automatically tokenized, lemmatized, and POS-tagged, and is currently undergoing manual correction.

The next phase involves creating an open-source sense repository for Croatian, which is being developed based on an existing pedagogical dictionary (Authors, 2025). The repository will be enriched through a combination of manual and automated methods, including the use of large language models (LLMs) to define missing senses. Since domain-specific terms and certain multiword expressions (MWEs) (Odijk, 2013) posed challenges for the tested translation platforms, a new evaluation task was conducted to assess the competence of LLMs in translating MWEs. The underlying hypothesis was that if an LLM could successfully translate MWEs from English into Croatian, it should also be capable of adequately identifying and defining their senses. Some studies have shown that LLMs perform particularly well in the semantic interpretation of MWEs (Gantar, 2024).

Each English sentence was automatically translated in a separate prompt using an

adapted pipeline for two large language models: ChatGPT-4o and the recently developed Slovene GaMS-9B-Instruct (https://huggingface.co/cjvt/GaMS-9B-Instruct). A preliminary evaluation was conducted on the first 200 sentences. As the translations generated by the GaMS-9B-Instruct model contained a significant number of Serbian lexical items (e.g., *fudbal*, *holandski napadač*, *spoljni stručnjaci* instead of nogomet 'football', *nizozemski napadač* 'Dutch striker', *vanjski stručnjaci* 'outside experts'), this set of translations was excluded from further evaluation. Five linguists then compared the ChatGPT-4o translations with the manually validated automatic translations, and marked differences.

This paper presents an analysis of the most common differences between the automatic translation of MWEs from English into Croatian by an LLM and the human validation of machine translation. ChatGPT-4o demonstrates a high level of proficiency in handling MWEs as opposed to its predecessors in this translation task. Differences between the compared translations include: a) wrong terminological equivalents (e.g., *medicinski uvjeti / medicinska stanja* 'medical conditions', *Bézierove površine / Bézierove plohe* 'Bézier surfaces'); b) differences at the morphosyntactic level (*Otto nagrada / nagrada Otto* 'Otto Award'; *riževi nemiri / rižini nemiri* 'rice protest'); c) English-influenced literal translations, mostly in verbal MWEs (*uzeti ime / dobiti ime* 'take its name', *častiti kao sveca / štovati kao sveca* 'honour as a saint'), d) the treatment of metaphorical MWEs (*pod protestom / u znak protesta* 'under protest', *proces se raspada / proces se urušava* 'the process breaks down'), and e) named entities, which is a challenge in other languages, too (Krstev et al., 2024). The provisional typology will be used in developing templates for defining MWEs in the sense repository for Croatian.

Keywords: multiword expressions; semantics; sense repository; translation evaluation

References

Authors. (2025).

Krstev, C., Stanković, R. & Marković, A. (2024). Towards the semantic annotation of sr-elexis corpus: Insights into multiword expressions and named entities. *Proceedings of Joint Workshop on Multiword Expressions and Universal Dependencies* (MWE-UD 2024).

Čibej, J. et al. (2025). Parallel sense-annotated corpus ELEXIS-WSD 1.2. *Slovenian language resource repository CLARIN.SI*, ISSN 2820-4042. Accessed at: http://hdl.handle.net/11356/2022.

Gantar, A. (2024). Formulisanje rečničkih definicija pomoću veštačke inteligencije na primeru slovenačkih frazeoloških jedinica. *Leksikografski susreti.* Marjanović, Saša (Ed.). Beograd: Filološki fakultet, pp. 151–158.

Martelli, F., Navigli, R., Krek, S., Kallas, J., Gantar, P., Koeva, S., Nimb, S., Pedersen, B., Olsen, S., Langemets, M., Koppel, K., Üksik, T., Dobrovoljc, K., Ureña Ruiz,

R., Sancho Sánchez, J.L., Lipp, V., Váradi, T., Győrffy, A., László, S. & Munda, T. (2021). Designing the ELEXIS Parallel Sense-annotated Dataset in 10 European Languages. *Electronic lexicography in the 21st century. Proceedings of the eLex 2021 conference*, pp. 377–395. Available at: https://elex.link/elex2021/wp-content/uploads/2021/08/eLex_2021_22_pp377-395.pdf.

Odijk, J. (2013). Identification and lexical representation of multiword expressions. *Essential Speech and Language Technology for Dutch. Results by the STEVIN-programme*. Spyns, Peter; Odijk, Jan (Eds.). Berlin - Heidelberg: Springer Berlin Heidelberg, pp. 201–217.

Up to No Good: Exploiting Word Embeddings for an Automatic Extraction of Candidates for a Lexicon of Slovene Taboo Language

Jaka Čibej

Centre for Language Resources and Technologies, University of Ljubljana
Faculty of Computer and Information Science, University of Ljubljana
Faculty of Arts, University of Ljubljana
Jožef Stefan Institute
E-mail: jaka.cibej@ff.uni-lj.si, jaka.cibej@ijs.si

Abstract

Lexicons of taboo language are useful language resources that can serve multiple purposes. In addition to their direct use to either automatically censor words deemed inappropriate for a given context (e.g. to help mitigate the problem of online hate speech), they can also help filter out materials not suitable for educational purposes (see Zingano Kuhn et al., 2022), games with a purpose (Arhar Holdt et al., 2021), training general language models (e.g. to remove pornographic content from training data). In addition, taboo language, particularly the section related to hate speech, needs to be well-documented in dictionaries as they are used as authoritative language resources (Gorjanc, 2005). Taboo language lexicons can also be useful for linguistic analyses and contrastive translation studies since swearing and taboo language are frequently culturally specific – see e.g. Klemenčič (2016) for a contrastive study of swearing in Slovene and Swedish; however, the study focused on a limited set of hand-picked expressions since no comprehensive list yet exists for Slovene, at least not in a machine-readable format.

What is included in existing Slovene language resources is either not openly accessible, is inaccurately represented (e.g. with *pejorative* as the only label, even though the context can be radically different in terms of intensity or taboeness: cf. *bedak* 'fool' vs. *peder* 'faggot'), or is limited in scope (*Thesaurus of Modern Slovene*; Krek et al., 2023), with material stemming mostly from corpora of standard Slovene, where the usage of offensive vocabulary is limited.

While similar lexicons have been compiled from existing language resources (e.g. van Huyssteen & Tiberius, 2023), we present an approach for constructing a list of Slovene taboo language candidates using the FastText embeddings trained on a number of Slovene corpora (including web-crawls). We first extract seed entries from the *Thesaurus of Modern Slovene 2.0* (Krek et al., 2023), which is part of the *Digital Dictionary Database of Slovene* (DDDS; Kosem et al., 2021). in which at least one of

the senses has been assigned a relevant label (hate speech, vulgar/coarse, expresses a negative attitude; see Arhar Holdt et al., 2022). We group them manually (e.g. religion-based, race-based, gender-based, homophobic slurs, words with sexual connotation), then use their embeddings (Terčon et al., 2023) and cross-compare them with other embeddings using cosine similarity to obtain a list of candidates for similar words.

We discuss the results of this extraction as well as the advantages (e.g. the detection of non-standard words or words that are rare in the corpus and might not be detected through a frequency-based approach) and disadvantages of this approach (e.g., it focuses on single-word expressions and is lexeme-focused instead of sense-focused). The resulting lexicon will be made available under an open-access license (CC BY-SA 4.0), also as part of the *Sloleks Morphological Lexicon of Slovene* (Čibej et al., 2022), which is part of the DDDS. The lexicon can provide a basis for a more detailed lexicographic analysis within DDDS, and the method can be applied to other languages.

Keywords: taboo language; automatic extraction; embeddings; corpora; Slovene

References

Arhar Holdt, Š., Logar, N., Pori, E. & Kosem, I. (2021). "Game of Words": Play the Game, Clean the Database. In Z. Gavriilidou, M. Mitsiaki, A. Fliatouras (eds.) *Lexicography for Inclusion: Proceedings of the 19th EURALEX International Congress*, 7-9 September 2021, Alexandroupolis, Vol. 1, pp. 41–49. Available at: https://www.euralex.org/elx_proceedings/Euralex2020-2021/EURALEX2020 2021_Vol1-p041-049.pdf.

Arhar Holdt, Š., Gantar, P., Kosem, I., Pori, E., Logar, N., Gorjanc, V. & Krek, S. (2022). Sovražno in grobo besedišče v odzivnem Slovarju sopomenk sodobne slovenščine. *Proceedings of the Conference on Language Technologies & Digital Humanities*. Ljubljana, 2022. Available at: https://nl.ijs.si/jtdh22/pdf/JTDH2022_ ArharHoldt-et-al_Sovrazno-in-grobo-besedisce-v-odzivnem-Slovarju-sopomenk-sodobne-slovenscine.pdf.

Čibej, J. et al. (2022). *Morphological lexicon Sloleks 3.0*, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042. Accessed at: http://hdl.handle.net/11356/1745.

Gorjanc, V. (2005). Neposredno in posredno žaljiv govor v jezikovnih priročnikih: diskurz slovarjev slovenskega jezika. *Družboslovne razprave* 21(48), pp. 197–209. Available at: http://dk.fdv.uni-lj.si/dr/dr48Gorjanc.PDF.

Klemenčič, I. (2016). *Fan, vad gör du?!: om svärande och svordomar i det svenska och slovenska språket*. BA Thesis. University of Ljubljana.

Kosem, I., Krek, S. & Gantar, P. (2021). Semantic data should no longer exist in isolation: the digital dictionary database of Slovenian. In Z. Gavriilidou, L. Mitits, S. Kiosses (eds.) *Proceedings of the XIX EURALEX International Congress: Lexicography for Inclusion*. Komotini: SynMorPhoSe Lab, Democritus University of Thrace, pp. 81–83.

Krek, S. et al. (2023). *Thesaurus of Modern Slovene 2.0*, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042. Accessed at: http://hdl.handle.net/11356/1916.

Terčon, L., Ljubešić, N. & Erjavec, T. (2023). *Word embeddings CLARIN.SI-embed.sl 2.0*, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042. Accessed at: http://hdl.handle.net/11356/1791.

Van Huyssteen, G.B., Tiberius, C. (2023). Towards a lexical database of Dutch taboo language. *2023: Electronic lexicography in the 21st century. Proceedings of eLex 2023 conference.* Available at: https://elex.link/ojs/index.php/elex/article/view/13.

Zingano Kuhn, T., Arhar Holdt, Š., Kosem, I., Tiberius, C., Koppel, K. & Zviel-Girshin, R. (2022). Data preparation in crowdsourcing for pedagogical purposes: the case of the CrowLL game. *Slovenščina 2.0: Empirične, aplikativne in interdisciplinarne raziskave* 10(2), pp. 62–100. Available at: https://doi.org/10.4312/slo2.0.2022.2.62-100.

Introducing DigiMet: a Psycholinguistic Database for Croatian Multi-Word Expressions

Jasmina Jelčić Čolakovac

University of Rijeka, Faculty of Maritime Studies
E-mail: jasmina.jelcic@uniri.hr

Abstract

The lack of normative resources for the Croatian language has incited the development of a novel resource that would not only compile normative data for Croatian but also focus on an underrepresented group of linguistic units – figurative multi-word (MWE) expressions. Thus, the creation of a normative database for figurative MWEs in Croatian is a significant step in the right direction that will address the gap in the availability of such tools for the Croatian language.

There are currently several normative databases available for Croatian single words such as the Croatian Psycholinguistic Database (Peti Stantić et al., 2021), psycholinguistic databases of affective norms and emotions (Ćoso et al., 2019; 2023), and the database of norms for non-adapted English words ENGRI CROWD (Bogunović et al., 2024). Given that all of the above sources contain normative data for individual words, a need arises to create a similar tool that would showcase norms for multi-word units. There is currently only one such database available; COMETA database (Citron et al. 2020) of affective and psycholinguistic norms for German conceptual metaphors is an open-access database featuring norms for emotional valence and arousal, imageability, and metaphoricity for conventional metaphors in both sentence and story contexts.

This is why the DigiMet database has been planned for development as a tool that will systematically catalog affective and lexico-semantic norms for Croatian metaphors along six different dimensions – 1. valence, 2. arousal, 3. concreteness, 4. imageability, 5. metaphoricity i 6. familiarity. The collection of norms will be carried out on a minimum sample of 500 native Croatian speakers using online distribution platforms such as SurveyMonkey. For this purpose, a combination of contrastive corpus research and manual data checking was carried out in the initial research phase. Using the MetaNet.HR database and corpus search in SketchEngine (SkE) (hrWaC 2.2, MaCoCu, enTenTen21), metaphors detected in Croatian (J1) and English (J2) and related MWEs were selected (verb-noun collocations were chosen as representative form of MWEs due to proven productivity in different languages). Lexical-semantic data on metaphorical MWEs was also extracted.

The DigiMet database, in its final form, will represent the first openly accessible

repository of metaphor norms for the Croatian language, which also represents the first database of affective and lexical-semantic data for Croatian multi-word expressions. This resource will enable further cross-linguistic comparisons and interdisciplinary experimental research.

Keywords: DigiMet database; figurative expressions; affective norms; Croatian metaphors; interdisciplinary research

Accelerating the lexicographic process with automatic methods and AI

Nathalie Norman[1], Nicolai Hartvig Sørensen[2], Jonas Jensen[2], Kirsten Appel[2], Sanni Nimb[2]

[1] University of Copenhagen, Njalsgade 80, DK-2300 S
[2]The Society for Language and Literature, Christians Brygge 1, DK-1219 K
E-mail: nhs@dsl.dk, naha@hum.ku.dk, jj@dsl.dk, ka@dsl.dk, sn@dsl.dk

Abstract

Writing dictionary entries is not only time-consuming but also an expensive process due to the highly specialized knowledge and experience required of the lexicographer. To facilitate the task of compiling the Danish monolingual dictionary DDO (ordnet.dk/ddo), we aim to establish an automatic assistant based on applied language technology (e.g. n-gram analysis, word embeddings, etc.) and generative AI. DDO contains 105,000 lemmas and is continuously updated with new lemmas twice a year. In this presentation, we focus on morphological and phonetic information in the dictionary, on synonyms and finally on an experiment with automatic writing of definitions.

The assistant, which we have named the Article Accelerator, automatically generates XML-tagged drafts of the subsections of a complete dictionary article in DDO. When the assistant gets a new word for the dictionary as input, it will automatically present suggestions for inflection, phonetic transcription, and synonyms. We assume that most new words in our case are compound nouns. In Danish, these are usually written together as a single word, and we therefore base the suggestions on a compound splitter. If the final part of the compound is already described in the dictionary, the assistant extracts the conjugation paradigms from the relevant entry or entries, and the user (i.e. the lexicographer) can then choose the appropriate one. Likewise, the assistant extracts the phonetic transcription for all subparts of a compound word that can be found in the dictionary. Lastly, synonyms are found by using both word embeddings and an LLM to get a list of synonym candidates. If a selected candidate already exists in the dictionary, the assistant can help create the necessary links and ID numbers.

The core of the Article Accelerator, however, is the module that generates suggestions for sense definitions based on existing definitions for semantically similar or related senses in the dictionary. These are found by combining compound splitting with a word embedding model. However, it is the user (i.e. the lexicographer) who selects the final list of senses, which are then included in the input to a generative model.

The goal is for the model to produce new definitions that reflect the style of the dictionary and require only minimal post-editing by the lexicographer. To find the optimal combination of prompt and generative model, we perform an experiment with fully edited but unpublished monosemous lemmas from DDO. We test two different prompts on three models (ChatGPT 4o, Claude 3.7 Sonnet, Llama 4 Scout) and manually compare the model's output with the definition written by a lexicographer.

The manual evaluation is carried out by two experienced lexicographers. This gives us knowledge about the quality of the automatic definitions and gives us the best conditions for choosing the ideal prompt and model.

Keywords: Generative models; AI assistant; automatic definition generation

Presenting verbal aspect data in a learner's dictionary: Devices and usage scenarios

Sarah Piepkorn

Hildesheim University, Universitätsplatz 1, 31141 Hildesheim, Germany
E-mail: piepkorns@uni-hildesheim.de

Abstract

A system of lexicographic presentational devices for data on verbal aspect has been developed that is aimed at providing advanced foreign language learners of English, German or Italian with data for individual verbs and their different readings. It is part of a monolingual, production-oriented electronic dictionary, the Phrase-based Active Dictionary (DiMuccio-Failla, 2025; DiMuccio-Failla & Giacomini, 2022).

Verbal aspect is understood here as the way in which speakers structure events and situations in language with regard to their boundaries (Sasse, 2002, p. 201). It is a conceptual category that is language-specific (Dessì Schmid, 2014), which means that providing data on verbal aspect can be beneficial for foreign language learners. Verbal aspect is expressed by the verb and its combination with linguistic devices, e. g. adverbials and tense, and it is tied to individual verb readings: Every verb reading has its characteristic set of 'aspectual properties' from a semantic as well as a syntactic point of view. For analysis, aspectual properties can be subsumed under more general aspectual classes (i. a. Vendler, 1957; Mourelatos, 1978; Croft, 2012).

The suggested system of presentational devices for verbal aspect consists of: 1) a visual representation of the aspectual class and corresponding semantic properties of the verb reading, 2) combinatorial options (adverbials, verbs and tense), 3) usage notes with explanations on semantic and/or syntactic particularities and 4) aids for disambiguating similar verb readings. The devices provide a range of data for the targeted user group of advanced language learners and are placed in different parts of the dictionary's article structure: The visual representations and combinatorial options are given alongside every verb reading. The usage notes are tied to the specific items the explanations refer to. The aids for disambiguating similar verb readings contain a link to their similar counterpart. Each type of device is associated with a symbol and the symbols are placed in the dictionary article as buttons to allow users to display the data on demand.

To illustrate the potential information gain for the target users, the presentational devices are demonstrated and related to usage situations from function theory (Tarp, 2008): text production, (the text production stage of) translation into the foreign language and the revision of existing texts. We describe how the presentational devices

cater to user needs in these situations and how they integrate with other microstructural items. The individual devices cater to different usage- and function-related user needs depending on the usage situation and user needs of a usage situation are covered by different devices. We exemplify the devices as well as different access routes within the dictionary, including aspect-class-based access via the above-mentioned visualisations.

Keywords: learner's lexicography; function theory; verbal aspect; actionality

References

Croft, W. (2012). *Verbs: Aspect and causal structure*. Oxford: Oxford University Press.

Dessì Schmid, S. (2014). *Aspektualität: Ein onomasiologisches Modell am Beispiel der romanischen Sprachen*. Berlin/Boston: De Gruyter. Available at: https://doi.org/10.1515/9783110334449.

DiMuccio-Failla, P.V. (2025). A theory for a usage-based cognitive lexicography. In L. Giacomini & V. Piunno (eds.) *Patterns of meaning in lexiography and lexicology*. Berlin/Boston: De Gruyter, pp. 19–90. Available at: https://doi.org/10.1515/9783111545943-003.

DiMuccio-Failla, P. V., & Giacomini, L. (2022). A proposed microstructure for a new kind of active learner's dictionary. *Lexicographica*, 38, pp. 475–499. Available at: https://doi.org/10.1515/lex-2022-0016.

Mourelatos, A.P.D. (1978). Events, processes, and states. *Linguistics and Philosophy*, 2(3), pp. 415–434. Available at: https://www.jstor.org/stable/25000995.

Sasse, H.-J. (2002). Recent activity in the theory of aspect: Accomplishments, achievements, or just non-progressive state? *Language Typology*, 6(2), pp. 199–271. Available at: https://doi.org/10.1515/lity.2002.007.

Tarp, S. (2008). *Lexicography in the borderland between knowledge and non-knowledge: General lexicographical theory with particular focus on learner's lexicography*. Berlin/New York: Max Niemeyer Verlag. Available at: https://doi.org/10.1515/9783484970434.

Vendler, Z. (1957). Verbs and times. *The Philosophical Review*, 66(2), pp. 143–160. Available at: https://doi.org/10.2307/2182371.

lexicographR: R infrastructure to develop and deploy digital dictionaries from scratch

Ligeia Lugli

Mangalam Research Center
E-mail: ligeialugli@hotmail.com

Abstract

This demo introduces lexicographR (citation withheld for anonymization), a prototype computer application aimed at facilitating the creation of digital dictionaries for scholars working in low-tech environments, where access to programming skills is severely hindered by lack of funding, institutional support and technical training. Based on recent user-surveys (Lugli 2024b), these scholars are typically domain experts or language teachers without formal training in lexicography and work on specialized dictionaries pertaining to their area of expertise. As such, they are often not aware of best practices and current methods in lexicography. Few use corpora and many have been writing their dictionaries in Word or Excel files, which makes it harder for them to automatically integrate new lexical data from corpora into their existing work. They typically struggle to deploy their lexicographic output as interactive online resources, and perceive existing free-of-charge digital dictionary development solutions, such as Lexonomy and Living Dictionaries (Daigneault and Anderson 2023; Měchura 2017), as insufficiently customisable for their highly specialized dictionaries and the specific needs of target audiences (Lugli 2024). The demo will first discuss the results of our user surveys and user-need identification process. It will then briefly discuss our development philosophy, which, given the ephemeral nature of interfaces and web-technologies, prioritizes lowering the costs and technical barrier to the creation of machine-readable and re-usable dictionary data over the development of digital interfaces. Still, to foster the dissemination of dictionary data among strata of the population who are less used to interacting with data directly, we have also provided a simple way to build flexible and lightweight interfaces to deploy dictionary data online as interactive digital dictionaries.

The core of the demo will consist of a demonstration of lexicographR's main functionalities, each of which is designed to assistance with a specific lexicographic task:

1. conversion of pre-existing dictionary data from Word, Excel, csv/tsv and FLEx, CoNLL-u and vrt/vert files into JSON.

2. processing corpus data from CoNLL-u, vrt/vert, csv/tsv, FLEx and plain text and extracting corpus frequencies nd distribution information for each dictionary headword

3. extracting collocations from the corpus for each dictionary headwords

4. extracting from the corpus for each dictionary headwords

5. creating data-visualizations for the information extracted from the corpus as well as for pre-existing dictionary data

6. designing a dictionary interface and generating the files necessary to publish the pre-existing dictionary data (potentially augmented with information extracted from the corpus and data-visualization) as either a Shiny app or a Quarto book.

7. converting the dictionary data published in the digital dictionary to JSON-LD for release in online data repositories, such as Zenodo or figshare.

The paper will conclude with an overview of some of the dictionaries that have been created using the lexicographR app.

Keywords: dictionary writing system; digital dictionary development tool;

lexicography software

References

Daigneault, A.L. & Anderson, G.D.S. (2023). Living Dictionaries: A Platform for Indigenous and Under-Resourced Languages. *Dictionaries: Journal of the Dictionary Society of North America,* 44(2), pp. 57–74. Available at: https://dx.doi.org/10.1353/dic.2023.a915065.

Lugli, L. (2024). Agile Lexicography: Rapid Dictionary Prototyping with R Shiny, with Examples from Projects on Sanskrit and Tibetan. In *Structuring Lexical Data and Digitising Dictionaries*. Leiden, The Netherlands: Brill. Chapter 6, pp. 121–149.

Lugli, L. (2024b). Democratizing Digital Lexicography: While Paper. Available at: https://figshare.com/collections/_/7207656.

Měchura, M.B. (2017). Introducing Lexonomy: an open-source dictionary writing and publishing system. In Electronic Lexicography in the 21st Century: Lexicography from Scratch. *Proceedings of the eLex 2017 conference*, Leiden.

Digitalization of Romanian dictionaries

Mititelu Catalin

"Iorgu Iordan - Alexandru Rosetti" Institute of Linguistics, Romanian Academy
E-mail: catalinmititelu@yahoo.com

Abstract

Recently, the digitization of resources of any type has become an increasingly discussed topic. In the linguistic field, lexicography is among the most influenced by this process, with digital dictionaries playing an essential role both for online consultation by specialists and for the automatic development of useful resources in natural language processing, as well as downstream applications.

The first dictionary automatically digitized by the "Iorgu Iordan - Alexandru Rosetti" Institute of Linguistics and made available to the public is the Etymological Dictionary of Romanian (https://delr.lingv.ro). It was parsed only shallowly, to make possible searches by the head word of the lexical entry, its variants and words from the same lexical family. It was developed rather as a proof of concept for the automatic parsing of the entries in dictionaries developed traditionally and originally meant only for printing.

The third edition of the Orthographic, Orthoepic and Morphological Dictionary of the Romanian Language (DOOM3) was produced by the Institute, initially also in printed format. Shortly after its paper format's launch on the market, the idea of making it accessible online to the general public and in a format that meets the current needs of users (i.e., quick access on mobile devices) led to its publication on the Internet (https://doom.lingv.ro), in a manner that allows for regular searches (by the title word), but also advanced ones (for example, by combining the various types of linguistic information represented in the dictionary: parts of speech, grammatical categories, language of origin, register, variants, etc.). The latter was made possible by the deeper parsing of its entries. Also, the entire theoretical apparatus that precedes the dictionary itself in the printed version, i.e. the Introductory Study, is also accessible online, which facilitates working with it, through the possibility of automatically searching its content for occurring words.

The online version is a more complex tool than the printed dictionary, because it has implemented a mechanism for suggesting the correct forms in the event that the user enters, in the search bar, a wrong word or forms that are no longer recommended/accepted by the norm.

Following the success among students, specialists, teachers and the large public of the digital edition of the Orthographic, Orthoepic and Morphological Dictionary, the

Institute invested effort in the digitalization of the new edition of the Romanian Language Dictionary (DLR). A new graphical interface has recently been created. For the moment, searches can only be made by the title word and are of several types: exact search, search with/without diacritics, search with prefixes or suffixes using the special characters * and ? (for example ab* or *tor, for prefixes and suffixes, respectively). The dictionary article contains several dynamic elements, especially regarding quotations, which are displayed compactly. Upon request, the user can see all quotations of a meaning or hide them completely for a synthetic view of the semantic tree (see https://dlr-test.lingv.ro/cautare/abandon). It is also possible to browse through the list of all words or download the list of words when searching with prefixes or suffixes.

In the future, we would like to add an advanced search that can be done according to criteria including: part of speech, register/usage, as well as consider other lexicographic resources to be made available online.

The method used to transpose the printable format into the online version is the same for all three dictionaries, despite the fact they have different structures.

Keywords: dictionary entry parsing; electronic dictionary; mobile devices

The Challenges of Syntactic Descriptions of Multiword Expressions in Electronic Lexicography

Verginica Barbu Mititelu[1], Voula Giouli[2], Gražina Korvel[3], Chaya Liebeskind[4], Irina Lobzhanidze[5], Rusudan Makhachashvili[6], Stella Markantonatou[7], Alexandra Markovic[8], Ivelina Stoyanova[9]

[1] Romanian Academy Research Institute for Artificial Intelligence, Bucharest
[2] Aristotle University of Thessaloniki, Thessaloniki
[3] Vilnius University, Vilnius
[4] Jerusalem College of Technology Jerusalem, Israel
[5] Ilia State University, Tbilisi
[6] Borys Grinchenko Kyiv Metropolitan University, Kyiv
[7] Institute for Language and Speech Processing and Archimedes/Athena RC, Athens
[8] Institute for Serbian Language, Belgrade
[9] Institute for Bulgarian Language, Bulgarian Academy of Sciences, Sofia
E-mail: vergi@racai.ro, pgiouli@del.auth.gr, grazina.korvel@mif.vu.lt, liebchaya@gmail.com, irina_lobzhanidze@iliauni.edu.ge, r.makhachashvili@kubg.edu.ua, marks@athenarc.gr, aleksandra.markovic@isj.sanu.ac.rs, iva@dcl.bas.bg

Abstract

In this paper, we provide a comprehensive overview of the way in which the morpho-syntactic properties of multiword expressions are represented in lexical resources to support Natural Language Processing downstream applications. Starting from an up-to-date and comprehensive overview of the existing lexica dedicated to multiword expressions and containing their syntactic description, we outline the current state of play in encoding syntactic information about multiword expressions (internal structure, argument structure, word order, discontinuity, verb alternations). We also discuss the relevance of the syntactic description of multiword expressions for several Natural Language Processing tasks. Our work contributes to the literature that fosters improvements in both the development and deployment of multiword expression lexica to ensure that they can support future Natural Language Processing innovations more effectively.

Keywords: multiword expression (MWE); lexica; morpho-syntactic description,

Natural Language Processing

References

Amaro, R., Giouli, V., Korvel, G., Lobzhanidze, I., Barbu Mititelu, V. & Valunaite Oleskeviciene, G. (2025). Perceptions on MWE lexicons use in NLP by the User Community: features, challenges and recommendations. Available at: https://unidive.lisn.upsaclay.fr/lib/exe/fetch.php?media=meetings:general_meetings:3rd_unidive_general_meeting:27_perceptions_on_mwe_lexicons.pdf. 3rd UniDive Workshop in Budapest.

Augustinus, L., Vandeghinste, V., Schuurman, I. & Van Eynde, F. (2013). Example-Based Treebank Querying with GrETEL–Now Also for Spoken Dutch. In S. Oepen, K. Hagen & J.B. Johannessen (eds.) *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*. Oslo, Norway: Linköping University Electronic Press, Sweden, pp. 423–428. Available at: https://aclanthology.org/W13-5638/.

Augustinus, L., Vandeghinste, V., Schuurman, I. & Van Eynde, F. (2017). GrETEL: A Tool for Example-Based Treebank Mining. In J. Odijk & A. van Hessen (eds.) *CLARIN in the Low Countries*, chapter 22. London, UK: Ubiquity, pp. 269–280. License: CC-BY 4.0.

Baldwin, T. & Kim, S.N. (2010). Multiword Expressions. In F.J. Damerau & N. Indurkhya (eds.) *Handbook of Natural Language Processing*. Chapman and Hall/CRC, 2 edition, pp. 267–292.

Barbu Mititelu, V., Giouli, V., Korvel, G., Liebeskind, C., Lobzhanidze, I., Makhachashvili, R., Markantonatou, S., Markovic, A. & Stoyanova, I. (2025). Survey on Lexical Resources Focused on Multiword Expressions for the Purposes of NLP. In A.K. Ojha, V. Giouli, V.B. Mititelu, M. Constant, G. Korvel, A.S. Doğruöz & A. Rademaker (eds.) *Proceedings of the 21st Workshop on Multiword Expressions (MWE 2025)*. Albuquerque, New Mexico, U.S.A.: Association for Computational Linguistics, pp. 41–57. Available at: https://aclanthology.org/2025.mwe-1.6/.

Bejček, E., Straňák, P. & Pecina, P. (2013). Syntactic Identification of Occurrences of Multiword Expressions in Text using a Lexicon with Dependency Structures. In V. Kordoni, C. Ramisch & A. Villavicencio (eds.) *Proceedings of the 9th Workshop on Multiword Expressions*. Atlanta, Georgia, USA: Association for Computational Linguistics, pp. 106–115. Available at: https://aclanthology.org/W13-1016/.

Borin, L., Forsberg, M. & Lyngfelt, B. (2013). Close encounters of the fifth kind: Some linguistic and computational aspects of the Swedish FrameNet++ project. *Veredas*, 17(1), pp. 28–43.

Boros, T., Pipa, S., Barbu Mititelu, V. & Tufis, D. (2017). A data-driven approach to verbal multiword expression detection. PARSEME Shared Task system description paper. In S. Markantonatou, C. Ramisch, A. Savary & V. Vincze

(eds.) *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*. Valencia, Spain: Association for Computational Linguistics, pp. 121–126. Available at: https://aclanthology.org/W17-1716/.

Campbell, R. (2004). Using linguistic principles to recover empty categories. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Chakrabarty, T., Saakyan, A., Ghosh, D. & Muresan, S. (2022). FLUTE: Figurative Language Understanding through Textual Explanations. In Y. Goldberg, Z. Kozareva & Y. Zhang (eds.) *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 7139–7159. Available at: https://aclanthology.org/2022.emnlp-main.481/.

Colson, J.P. (2020). HMSid and HMSid2 at PARSEME Shared Task 2020: Computational Corpus Linguistics and unseen-in-training MWEs. In S. Markantonatou, J. McCrae, J. Mitrović, C. Tiberius, C. Ramisch, A. Vaidya, P. Osenova & A. Savary (eds.) *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*. online: Association for Computational Linguistics, pp. 119–123. Available at: https://aclanthology.org/2020.mwe-1.15/.

Cook, P., Fazly, A. & Stevenson, S. (2007). Pulling their Weight: Exploiting Syntactic Forms for the Automatic Identification of Idiomatic Expressions in Context. In N. Gregoire, S. Evert & S.N. Kim (eds.) *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*. Prague, Czech Republic: Association for Computational Linguistics, pp. 41–48. Available at: https://aclanthology.org/W07-1106/.

Cordeiro, S., Ramisch, C. & Villavicencio, A. (2016). mwetoolkit+ sem: Integrating word embeddings in the mwetoolkit for semantic MWE processing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 1221–1225.

Duran, M.S. & Ramisch, C. (2011). How do you feel? investigating lexical-syntactic patterns in sentiment expression. In *Proceedings of corpus linguistics*.

Ebrahim, S., Hegazy, D., Mostafa, M.G.H.M. & El-Beltagy, S.R. (2017). Detecting and integrating multiword expression into English-Arabic statistical machine translation. *Procedia Computer Science*, 117, pp. 111–118.

Giouli, V. & Barbu Mititelu, V. (2024). *Multiword expressions in lexical resources: Linguistic, lexicographic, and computational perspectives*. Language Science Press.

Ide, Y., Tanner, J., Nohejl, A., Hoffman, J., Vasselli, J., Kamigaito, H. & Watanabe, T. (2025). CoAM: Corpus of All-Type Multiword Expressions. In W. Che, J. Nabende, E. Shutova & M.T. Pilehvar (eds.) *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vienna, Austria: Association for Computational Linguistics, pp. 27004–27021. Available at: https://aclanthology.org/2025.acl-long.1311/.

Iñurrieta, U., Aduriz, I., Díaz de Ilarraza, A., Labaka, G. & Sarasola, K. (2018). Konbitzul: an MWE-specific database for Spanish-Basque. In N. Calzolari, K.

Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis & T. Tokunaga (eds.) *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA). Available at: https://aclanthology.org/L18-1397/.

Johnson, M. (2002). A simple pattern-matching algorithm for recovering empty nodes and their antecedents. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Levin, B. (1993). *English Verb Classes and Alternations*. University of Chicago Press.

Lion-Bouton, A., Savary, A. & Antoine, J.Y. (2023). A MWE lexicon formalism optimised for observational adequacy. In A. Bhatia, K. Evang, M. Garcia, V. Giouli, L. Han & S. Taslimipoor (eds.) *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*. Dubrovnik, Croatia: Association for Computational Linguistics, pp. 121–130. Available at: https://aclanthology.org/2023.mwe-1.16/.

Losnegaard, G.S., Sangati, F., Escartín, C.P., Savary, A., Bargmann, S. & Monti, J. (2016). PARSEME Survey on MWE Resources. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk & S. Piperidis (eds.) *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Portorož, Slovenia: European Language Resources Association (ELRA), pp. 2299–2306. Available at: https://aclanthology.org/L16-1364.

Maziarz, M., Grabowski, Ł., Piotrowski, T., Rudnicka, E. & Piasecki, M. (2023). Lexicalisation of Polish and English word combinations: an empirical study. *Poznan Studies in Contemporary Linguistics*, 59(2), pp. 381–406.

Miletić, F. & Walde, S.S.i. (2024). Semantics of Multiword Expressions in Transformer-Based Models: A Survey. *Transactions of the Association for Computational Linguistics*, 12, pp. 593–612. Available at: https://aclanthology.org/2024.tacl-1.33/.

Mohamed, N.H., Savary, A., Khelil, C.B., Antoine, J.Y., Keskes, I. & Belguith, L.H. (2024). Lexicons Gain the Upper Hand in Arabic MWE Identification. In *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWEUD)@LREC-COLING 2024*. pp. 88–97.

Moreau, E., Alsulaimani, A., Maldonado, A. & Vogel, C. (2018). CRF-Seq and CRFDepTree at PARSEME Shared Task 2018: Detecting Verbal MWEs using Sequential and Dependency-Based Approaches. In A. Savary, C. Ramisch, J.D. Hwang, N. Schneider, M. Andresen, S. Pradhan & M.R.L. Petruck (eds.) *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWECxG-2018)*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 241–247. Available at: https://aclanthology.org/W18-4926/.

Muller, I., Mamede, N. & Baptista, J. (2024). Hurdles in Parsing Multi-word Adverbs: Examples from Portuguese. In P. Gamallo, D. Claro, A. Teixeira, L. Real, M.

Garcia, H.G. Oliveira & R. Amaro (eds.) *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1.* Santiago de Compostela, Galicia/Spain: Association for Computational Linguistics, pp. 177–186. Available at: https://aclanthology.org/2024.propor-1.18.

Nivre, J., Hall, J. & Nilsson, J. (2006). MaltParser: A Data-Driven Parser-Generator for Dependency Parsing. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. Genoa, Italy: European Language Resources Association (ELRA).

Nivre, J. & Nilsson, J. (2005). Pseudo-Projective Dependency Parsing. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*. Ann Arbor, Michigan: Association for Computational Linguistics, pp. 99–106.

Odijk, J., Kroon, M., Spoel, S., Bonfil, B. & Baarda, T. (2024). *Querying for multiword expressions in large Dutch text corpora*. Phraseology and Multiword Expressions. Language Science Press, pp. 229–267. Publisher Copyright: © 2024, the authors. All rights reserved.

Phelps, D., Pickard, T.M.R., Mi, M., Gow-Smith, E. & Villavicencio, A. (2024). Sign of the Times: Evaluating the use of Large Language Models for Idiomaticity Detection. In A. Bhatia, G. Bouma, A.S. Doğruöz, K. Evang, M. Garcia, V. Giouli, L. Han, J. Nivre & A. Rademaker (eds.) *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024*. Torino, Italia: ELRA and ICCL, pp. 178–187. Available at: https://aclanthology.org/2024.mwe-1.22.

Ramisch, C. (2015). Multiword Expressions Acquisition: A Generic and Open Framework.

Ramisch, C., Cordeiro, S.R., Savary, A., Vincze, V., Barbu Mititelu, V., Bhatia, A., Buljan, M., Candito, M., Gantar, P., Giouli, V., Güngör, T., Hawwari, A., Iñurrieta, U., Kovalevskaitʹe, J., Krek, S., Lichte, T., Liebeskind, C., Monti, J., Parra Escartín, C., QasemiZadeh, B., Ramisch, R., Schneider, N., Stoyanova, I., Vaidya, A. & Walsh, A. (2018). Edition 1.1 of the PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions. In A. Savary, C. Ramisch, J.D. Hwang, N. Schneider, M. Andresen, S. Pradhan & M.R.L. Petruck (eds.) *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 222–240. Available at: https://aclanthology.org/W18-4925/.

Ramisch, C., Savary, A., Guillaume, B., Waszczuk, J., Candito, M., Vaidya, A., Barbu Mititelu, V., Bhatia, A., Iñurrieta, U., Giouli, V., Güngör, T., Jiang, M., Lichte, T., Liebeskind, C., Monti, J., Ramisch, R., Stymne, S., Walsh, A. & Xu, H. (2020). Edition 1.2 of the PARSEME Shared Task on Semi-supervised Identification of Verbal Multiword Expressions. In S. Markantonatou, J. McCrae, J. Mitrović, C. Tiberius, C. Ramisch, A. Vaidya, P. Osenova & A. Savary (eds.) *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*. online:

Association for Computational Linguistics, pp. 107–118. Available at: https://aclanthology.org/2020.mwe-1.14/.

Ramisch, C., Villavicencio, A. & Boitet, C. (2010). Multiword expressions in the wild? the mwetoolkit comes in handy. In *Coling 2010: Demonstrations*. pp. 57–60.

Rikters, M. & Bojar, O. (2017). Paying Attention to Multi-Word Expressions in Neural Machine Translation. In *Proceedings of Machine Translation Summit XVI: Research Track,* pp. 86–95.

Rohanian, O., Taslimipoor, S., Kouchaki, S., Ha, L.A. & Mitkov, R. (2019). Bridging the Gap: Attending to Discontinuity in Identification of Multiword Expressions. In J. Burstein, C. Doran & T. 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)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 2692–2698. Available at: https://aclanthology.org/N19-1275/.

Rosen, V., Losnegaard, G.S., De Smedt, K., Bejcek, E., Savary, A., Przepiorkowski, A., Osenova, P. & Barbu Mititelu, V. (2015). A survey of multiword expressions in treebanks. In *Proceedings of the Treebanks and Linguistic Theories conference (TLT - 2015)*. Warsaw, Poland, pp. 179–193. Available at: https://ufal.mff.cuni.cz/biblio/attachments/2015-bejcek-m9168853603341436530.pdf.

Sag, Ivan A. & Baldwin, T., Bond, F., Copestake, A. & Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP. In *Lecture Notes in Computer Science. Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, volume 2276. Springer, pp. 189–206.

Sánchez Cárdenas, B. & Ramisch, C. (2019). Eliciting specialized frames from corpora using argument-structure extraction techniques. *Terminology*, 25(1), pp. 1–31.

Savary, A., Cordeiro, S. & Ramisch, C. (2019). Without lexicons, multiword expression identification will never fly: A position statement. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*. Florence, pp. 79–91.

Savary, A., Ramisch, C., Cordeiro, S., Sangati, F., Vincze, V., QasemiZadeh, B., Candito, M., Cap, F., Giouli, V., Stoyanova, I. & Doucet, A. (2017). The PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions. In S. Markantonatou, C. Ramisch, A. Savary & V. Vincze (eds.) *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*. Valencia, Spain: Association for Computational Linguistics, pp. 31–47. Available at: https://aclanthology.org/W17-1704/.

Scholivet, M., Ramisch, C. & Cordeiro, S. (2018). Sequence models and lexical resources for MWE identification in French. In *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, volume 2. Language Science Press, pp. 263.

Shayegh, B., Wen, Y. & Mou, L. (2024). Tree-Averaging Algorithms for Ensemble-Based Unsupervised Discontinuous Constituency Parsing. 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)*. Bangkok, Thailand: Association for Computational Linguistics, pp. 15135–15156. Available at: https://aclanthology.org/2024.acl-long.808/.

Skoumalová, H. & Kopřivová, M. (2024). *Multiword expressions in lexical resources: Linguistic, lexicographic, and computational perspectives*, chapter LEMUR: A lexicon of Czech multiword expressions. Language Science Press.

Tanner, J. & Hoffman, J. (2023). MWE as WSD: Solving Multiword Expression Identification with Word Sense Disambiguation. In H. Bouamor, J. Pino & K. Bali (eds.) *Findings of the Association for Computational Linguistics: EMNLP 2023*. Singapore: Association for Computational Linguistics, pp. 181–193. Available at: https://aclanthology.org/2023.findings-emnlp.14/.

Tayyar Madabushi, H., Gow-Smith, E., Scarton, C. & Villavicencio, A. (2021). AStitchIn-LanguageModels: Dataset and Methods for the Exploration of Idiomaticity in Pre-Trained Language Models. In M.F. Moens, X. Huang, L. Specia & S.W.t. Yih (eds.) *Findings of the Association for Computational Linguistics: EMNLP 2021*. Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 3464–3477. Available at: https://aclanthology.org/2021.findings-emnlp.294/.

Urešová, Z., Štěpánek, J., Hajič, J., Panevova, J. & Mikulová, M. (2014). PDTVallex: Czech Valency lexicon linked to treebanks. Available at: http://hdl.handle.net/11858/00-097C-0000-0023-4338-F. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).

Vondřička, P. (2019). Design of a multiword expressions database. *The Prague Bulletin of Mathematical Linguistics*, 112(1), pp. 83–101.

Waszczuk, J. (2018). TRAVERSAL at PARSEME Shared Task 2018: Identification of Verbal Multiword Expressions Using a Discriminative Tree-Structured Model. In A. Savary, C. Ramisch, J.D. Hwang, N. Schneider, M. Andresen, S. Pradhan & M.R.L. Petruck (eds.) *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 275–282. Available at: https://aclanthology.org/W18-4931/.

Zabokrtský, Z. & Lopatková, M. (2007). Valency Information in VALLEX 2.0: Logical Structure of the Lexicon. *Prague Bull. Math. Linguistics*, 87, pp. 41–60. Available at: http://ufal.mff.cuni.cz/pbml/87/zabokrtsky-lopatkova.pdf.

Zagatti, F., de Lima Medeiros, P.A., da Cunha Soares, E., dos Santos Silva, L.N., Ramisch, C. & Real, L. (2022). mwetoolkit-lib: Adaptation of the mwetoolkit as a Python Library and an Application to MWE-based Document Clustering. In *Proceedings of the 18th Workshop on Multiword Expressions@ LREC2022,* pp. 112–117.

Zhang, X., Bouma, G. & Bos, J. (2025). Neural Semantic Parsing with Extremely Rich Symbolic Meaning Representations. *Computational Linguistics*, 51(1), pp. 235–274. Available at: https://doi.org/10.1162/coli_a_00542.

https://direct.mit.edu/coli/article-pdf/. 51/1/235/2483888/coli_a_00542.pdf.

The dictionary of pluricentric Portuguese project: theoretical aspects

Tanara Zingano Kuhn

University of Coimbra
E-mail: tanarazingano@outlook.com

Abstract

The dictionary of pluricentric Portuguese project, which is at its initial stage at (University of Coimbra), aims at providing a free, online dictionary that describes Portuguese as it is used in several territories around the globe. The purpose of this poster is to present theoretical questions that need to be answered to guide the methodological decisions for the creation of this dictionary, bearing in mind our alignment with the idea of "socially responsible lexicography" (Calañas Continente & Domínguez Vázquez, 2023) and the socio-political-cultural complexities inherent to the Portuguese language area. From an official status viewpoint, Portuguese is used in nine countries and one territory. Nevertheless, the functional status of the language varies significantly in these regions, ranging from its status as the mother tongue of the majority of the population (Brazil, Sao Tome and Principe, Portugal), to its role as the predominant vehicular language, typically as a second language (Angola, Mozambique), to its status as a minority language (Cabo Verde, Guinea-Bissau, Timor-Leste), to the point of its virtual non-use (Guinea Equatorial, Macao). As to language standards, Brazil and Portugal have been traditionally considered norm-setting centres, having fully-fledged standardizing and codifying instruments such as dictionaries and grammars, with the European variety being adopted as the norm in the other countries. However, this bicentric view has been challenged by researchers who have shown that local varieties of Portuguese have been emerging in other countries. In addition, there is a growing demand in society for the recognition of these varieties as valid and legitimate as the dominant varieties, with the compilation of a Dictionary of Mozambican Portuguese currently underway (see Machungo & Firmino, 2022). This highlights the complex relationship between language, power, and identity. These complex socio-political-cultural contexts of all these multilingual territories, together with our ideological position to counter what Rizzo (2019: 287) has identified as "homogenizing tendencies in certain language policies that seek to impose a dominant reality", make the production of a dictionary of pluricentric Portuguese a highly challenging undertaking. One of the greatest challenges is the fact that, in territories where Portuguese was introduced as a result of colonisation, the dominant exonormative view of the language leads to a significant gap between how language is used on a daily basis and the use imposed by the school and other language regulators. This has several consequences to our lexicographic project, starting with the

establishment of what definition of norm is suitable to our project, which in turn will support decisions regarding the corpus to be used as a source. Considering all that in this dictionary project means that prior theoretical research must be carried out in order to inform the decision-making process regarding corpus compilation, headword candidate list, entry configuration, entry microstructure, to name but a few. In this poster, we will position ourselves in terms of theoretical references, present crucial questions for the making of the dictionary, and share tentative answers. We hope this paper will promote exchange of knowledge and experience with peer lexicographers facing similar challenges in their projects, as well as encourage reflection on the political role of lexicography (Crowley, 1999).

Keywords: Pluricentric Portuguese; Sociolinguistics; Dictionary-making

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.

Albuquerque, D. (2024). *A língua portuguesa falada em Timor-Leste: Um estudo ecolinguístico* [Portuguese Language Spoken in Timor-Leste: An ecolinguistic research]. PCL Press.

Alexandre, N. & Gonçalves, R. (2018). Language contact and variation in Cape Verde and São Tomé and Príncipe. In L. Álvarez López, P. Gonçalves, & J. Ornelas de Avelar (Eds.), *The Portuguese Language Continuum in Africa and Brazil,* pp. 237–265. John Benjamins Publishing Company.

Álvarez López, L., Gonçalves, P. & Ornelas de Avelar, J. (eds.) (2018). *The Portuguese Language Continuum in Africa and Brazil*. John Benjamins Publishing Company.

Bouchard, M.-E. (2017). *Linguistic variation and change in the Portuguese of São Tomé* (PhD thesis). New York University.

Calañas Continente, J.A. & Domínguez Vázquez, M.J. (2023). About Sustainable and Socially Responsible Lexicography: State of the Art. *Quaderns de Filologia: Estudis Lingüístics*, XXVIII, pp. 9–19. Available at: doi: 10.7203/QF.28.27659.

Crowley, T. (1999) The Socially Responsible Lexicographer in Oceania. *Journal of Multilingual and Multicultural Development*, 20(1), pp. 1–12. Available at: doi:10.1080/01434639908666366.

Faraco, C.A. (2016). *História sociolinguística da língua portuguesa* [Sociolinguistic history of the Portuguese Language]. 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.

Machungo, I. & Firmino, G. (2022). Variedades nacionais de línguas pluricêntricas: O caso do português de Moçambique [National varieties of pluricentric languages: The case of Mozambican Portuguese]. *Platô*, 5(9), pp. 28–45.

Pinto, P.F. & Melo-Pfeiffer, S. (eds.). (2018). *Políticas linguísticas em Português* [Language policies in Portuguese]. Lidel.

Rizzo, M.F. (2019). Discusiones actuales en torno a la lusofonía: panorama de los estudios sobre política internacional del portugues [Current discussions on Lusophony: an overview of the international policy studies of Portuguese]. *Trabalhos em Linguística Aplicada*, 58(1), pp. 287–312.