

LexiKonf 2025

1ST INTERNATIONAL CONFERENCE ON LEXICOLOGY AND LEXICOGRAPHY

BOOK OF ABSTRACTS

29th September – 1st October 2025 Budapest, Hungary

Edited by Júlia Ballagó and Veronika Lipp

ELTE Research Centre for Linguistics
Budapest
2025

LexiKonf 2025

1ST INTERNATIONAL CONFERENCE ON LEXICOLOGY AND LEXICOGRAPHY

BOOK OF ABSTRACTS

29th September – 1st October 2025 Budapest, Hungary

Edited by Júlia Ballagó and Veronika Lipp

ELTE Research Centre for Linguistics
Budapest
2025

Programme Committee

Programme Chairs:

Veronika Lipp, PhD

Júlia Ballagó, PhD

Members of the Programme Committee:

Enikő Héja, PhD

Bálint Sass, PhD

Réka Szabó, PhD

Éva Dömötör, PhD

Ágnes Kalivoda, PhD

Dóra Mária Tamás, PhD

Katalin P. Márkus, PhD

Kata Heller, PhD student

Judit Bakó, PhD student

Editorial assistants:

Dávid Timár

Zoltán Tarr

ISBN: 978-963-489-844-3

DOI: <https://doi.org/10.18135/LexiKonf2025>

© 2025

ELTE Research Centre for Linguistics, Budapest

Contents

Abstracts of Plenary Lectures

Sven Tarp: Do Lexicography a Favour: Stop Focusing on Dictionaries.....	7
Iztok Kosem: Common Sense(s) in Slovene Lexicography: Building the Digital Dictionary Database.....	8
Ana Salgado: A Decade of Lexicographic Innovation at the Lisbon Academy of Sciences: Still from Paper to Digital Platform – and Then AI Came.....	9

Abstracts of Papers

Salvatore Arcidiacono & Marco Venuti: Lexicad: A Framework for Digital Lexicography and Early Italian Language Resources.....	10
William Rolston Ashford: Building a better Word Bank: Organising lexicographic evidence at Dictionaries of the Scots Language	12
Katarína Balleková, Peter Malčovský & Mária Mikolajová: Online popularization of research on Slovak dialects – Digital Atlas of the Slovak Language (2024, 1.0)	13
Vladimír Benko: Needles in a Haystack? Looking for Neologisms in a Web-Crawled Corpus	15
Attila Benő, Katalin P. Márkus & Tibor M. Pintér: A Hungarian Dictionary in the Online Era: Termini Online Hungarian–Hungarian Dictionary and Database	17
Wiebke Blanck: Towards the pragmatic item swearword and its (current and future) role in general-language online dictionaries of German	18
Edmond Cane: A Construction-based model for language processing	19
István Csernicskó & Krisztián Váradi: The Role of Regional Dialects in the Development of an Educational Terminology Database and the Termini Hungarian–Hungarian Dictionary	21
Sara de Piniés de la Cuesta & Javier Serrano Peinado: Identifying Collocations in Oceanographic Lexicology: A Focus on Temperature, Pressure, and Salinity	22
Csilla Ilona Dér: Interjection-discourse markers in 19th–20th century Hungarian dictionaries, grammars and databases	23
Fazakas Emese & Zsemlyei Borbála: The necessity of an index volume to the Historical Dictionary of the Hungarian Language in Transylvania	25

Márta Fischer & Dóra Mária Tamás: Educational Terminology in Hungary and the Hungarian-speaking Communities of the Neighbouring Countries: Strategic Issues, Dilemmas and Solutions	26
Emmanuel Sylvain Fomat: A lexico-semantic analysis of neologisms in the Anglophone crisis in Cameroon	27
Péter Gaál: Dictionary use habits of teachers from the Prekmurje region - A case study from Slovenia	28
Radovan Garabík & Agáta Karčová: Analyzing grammatical anomalies in lexical data for fun and profit	29
Luisa Giacomà: The strange case of the lexicographic paratext	31
Anna Grzeszak: On unconventional multilingual dictionaries (17th–21st century)	32
Zita Hollós: Busy as a bee! From a corpus-guided collocations dictionary to a learner's dictionary portal for German	33
Jakob Horsch: The closer you look, the more interesting it gets: A Construction Grammar-based analysis of the Slovak Comparative Correlative	35
Lana Hudeček & Milica Mihaljević: The Croatian Web Dictionary – Mrežnik – from corpora to dictionary	37
Kathryn M. Hudson: Re-Rooting the Self: Historical Lexicography and Identity in Contemporary Mesoamerica	39
Katarzyna Kryńska: Implementing Greek into OCR Model of Multilingual Old Dictionary	40
Anna Lehocki-Samardžić & Szilvia Szoták: Examining the educational terms of the Termini Hungarian-Hungarian dictionary	42
Miguel Angel Lopez, Elena Andreeva & Larissa Ferreira: Ausgewanderte Wörter: An online Dictionary of German Emigrant Words	44
Daniela Majchráková: Compiling the Slovak Dictionary of Adverbial Collocations	45
Magdalena Majdak: The Database of Historical Polish Lexicons: Concept and Lexicographic Challenges	46
Susana Duarte Martins: Educational applications of general language dictionaries in advanced language proficiency	47

Maja Matijević: Autohyponymy and Automeronymy in the Croatian Language	49
Nathalie Mederake & Nico Urbach: Towards a broader view of regional dialects. Cross-linguistic connections in dialect dictionaries	50
Peter Meyer: Walking the graph: Finding etymological information in a web platform on German loanwords in other languages	51
Ana Mihaljević & Josip Mihaljević: The second phase of the development of the online version of the Dictionary of the Croatian Redaction of Church Slavonic – New challenges.....	52
Sylvain Monnoukoun: Research into Dictionary use: User Behaviors and Preferences in Benin	53
Perina Vukša Nahod & Bruno Nahod: From apron to fork – a new approach to Croatian dialectological dictionary	55
Barbara Patella: Digitize dictionaries using XML-TEI: a vademecum of methodologies and applications (based on non-alphabetic resources)	56
Marta Petrak & Ivana Franić: FraCroVal: a contribution to contrastive studies of verbal valency	61
Dóra Pődör: The First Irish-Hungarian / Hungarian-Irish Learner's Dictionary	62
Rui Qiao: Exploring Chengduhua through the Lexicographical Works of Missionaries (19th-20th Century)	63
Iasmin Valéria Miranda Rabelo, Adriana Silvina Pagano & Maucha Andrade Gamonal: Disability and Assistive Technology: Modeling Accessibility Domain Frames on FrameNet Brasil	65
Lydia Risberg, Eleri Aedmaa, Maria Tuulik, Margit Langemets, Ene Vainik, Esta Prangel, Kristina Koppel & Hanna Pook: Who decides what's informal? Reducing subjectivity in dictionary labeling with corpora and LLMs	66
Ewa Rodek: Training OCR Models for Historical Multilingual Dictionaries	68
Manana Rusieshvili-Cartledge & Marine Makhatadze: Exploring Sensory Lexis: Typological and Structural Perspectives of the Georgian-English Thematic Dictionary	70
Ersilia Russo: Phraseology in the Vocabolario del Fiorentino Contemporaneo ('Vocabulary of contemporary Florentine')	72

John M. Ryan & Víctor Parra-Guinaldo: A Lexicographic Approach to the Classification of Relexified Diminutives in the Romance Languages: Phase III – Neapolitan	74
Bálint Sass: Morphological dependency trees for representing constructions ..	76
Olívia Seidl-Péché: Terminological and Lexicological Challenges in Neural Machine Translation: A Case Study of EU Press Releases	78
Kganathi Shaku & Mmagonkahloleng Brudence Makau: Using digital platforms for data collection in lexicography: Mzansi Taal dictionary as case of analysis	80
Hindrik Sijens & Johan van der Zwaag: Towards a new historical dictionary of Frisian ..	81
Sven-Erik Soosaar, Madis Jürviste & Tiina Paet: Large Language Models as Tools for Historical Lexicography: Automating Homonym Detection.	82
Stefania Spina, Irene Fioravanti, Fabio Zanda, Luciana Forti, Damiano Perri & Osvaldo Gervasi: A learner dictionary of Italian collocations: proficiency-based attribution and AI-generated definitions.	84
Mónika Varga: From 'shape' to 'sort of' (from hedge to dress): a historical corpus study of szabású	86
Teresa Fuentes & Yuliia Vasik: Monolingual Lexicography and Visual Impairment: Parameters for More Accessible Dictionaries	87
Yelena Yerznkyan: Words with Pragmatic Loading: Main Principles of Dictionary Entry Development	88
Carlo Zoli: Bridging Dialect Lexicography and Geolinguistic Atlases: the Smallcodes Approach.....	89
Krisztina-Mária Sárosi-Márdírosz & Krisztina Sófalvi: Hungarian-Romanian terminological matches in the Carpathian Basin educational terminology database	91
Franck Zumstein: A comparative study of authoritative 18th century pronunciation dictionaries of the English language	92

Sven Tarp
Aarhus University, Denmark
st@cc.au.dk

Do Lexicography a Favour: Stop Focusing on Dictionaries

This paper, inspired by Al-Kasimi (1977), rests on two core assumptions: first, lexicographic products are historically and culturally contingent—shaped by the technological conditions and communicative needs of their time; second, current developments in generative AI are not merely enhancing dictionary compilation methods, but giving rise to entirely new lexicographic products designed to meet evolving user needs. Consequently, the traditional dictionary is losing its centrality as users increasingly seek alternative forms of linguistic assistance. Lexicographers must respond to this paradigm shift.

Building on Tarp and Gouws' (2023) redefinition of lexicography into two branches—dictionography and glossography—this paper explores the potential of the latter. Glossography encompasses both the historical glosses that predate dictionaries (Benati & Händl 2019) and the novel digital glosses emerging in today's AI-driven landscape (Gouws & Tarp 2024).

The focus is on three types of digital glosses developed within an AI-supported Writing Assistant designed to help L2 learners. These glosses are triggered by grammatical errors in user-generated texts and are presented as: (1) a brief suggestion to correct the error, (2) a short gloss explaining the nature of the mistake, and (3) an extended gloss detailing the relevant grammatical rule. These components do not reference specific lemmata but are problem-oriented, offering a dynamic alternative to traditional dictionary-based grammatical notes.

The paper introduces a conceptual framework based on Hegelian categories: the learner's mistake (individual), the short gloss (particular), and the long gloss (universal). Unlike Hegel's model, this system begins with the individual instance and ascends to the universal via the particular—thus supporting a transition from incidental to intentional learning.

In conclusion, this reverse-grammar model offers a more natural path to grammatical competence, grounded in the learner's own output and errors. Rather than applying pre-learned rules, learners encounter grammar as an embedded, contextualised, and personalised resource—suggesting that the future of lexicography may lie not in the dictionary, but in interactive, AI-powered glossographic solutions.

References:

- Al-Kasimi, A. (1977) *Linguistic and Bilingual Dictionaries*. Leiden: EJ Brill Publishers.
- Gouws, R.H. & Tarp, S. (2024). Despite Current Challenges: Lexicography has a bright future. *Lexicographica*, 40, 187-212.
- Tarp, S. & Gouws, R.H. (2023) A Necessary Redefinition of Lexicography in the Digital Age: Glossography, Dictionography and the Implications for the Future. *Lexikos*, 33, 425-447.

Iztok Kosem
University of Ljubljana, Slovenia
iztok.kosem@fri.uni-lj.si

Common Sense(s) in Slovene Lexicography: Building the Digital Dictionary Database

In this paper, I will present the recent lexicographic activities at the Centre for Language Resources and Technologies, University of Ljubljana, with a focus on the development of the Digital Dictionary Database of Slovene. This ongoing process has proved highly challenging, as it involves merging heterogeneous resources and datasets, deciding how to link external materials such as crowdsourced data, and establishing criteria for a shared repository of senses and concepts. I will illustrate these issues with selected examples of problems and solutions. While the fact that many resources are being compiled from scratch has been advantageous, it has also created problems—notably the need for subsequent cleaning and lemma labelling, since some resources were generated through post-editing approaches. In parallel, the migration from the external dictionary platform Lexonomy to an in-house database editor is nearing completion; this transition brings clear advantages but also entails additional training for lexicographers and a reconceptualization of lexicographic workflows. With the emergence of large language models, we have also been exploring and testing different options for their integration into the lexicographic process. Finally, I will present several cases demonstrating how the database is already being applied in non-lexicographic projects.

Ana Salgado

Academia das Ciências de Lisboa, Portugal

ana.salgado@acad-ciencias.pt

A Decade of Lexicographic Innovation at the Lisbon Academy of Sciences: Still from Paper to Digital Platform – and Then AI Came

This talk explores recent developments in the Portuguese Academy dictionary project within the context of digital transition and artificial intelligence. Drawing on ten years of work at the Lisbon Academy of Sciences, the presentation offers a first-hand account of the challenges and innovations involved in bringing the *Dicionário da Língua Portuguesa* into the digital age. It examines the shift from traditional editorial practices to corpus-driven methodologies, the redefinition of microstructural components, and the establishment of a sustainable digital platform. Special attention will be given to the recent integration of tools and methods into the lexicographic workflow, including criteria and strategies for updating content in response to linguistic evolution, shifting social trends, user log analysis, semi-automatic neologism detection, and language policies. By reflecting on institutional constraints, technological advances, and evolving user expectations, this talk invites a broader discussion on how Academy dictionaries can remain scientifically rigorous, inclusive, and socially relevant.

Salvatore Arcidiacono, Marco Venuti
Università degli Studi di Catania, Italy
salvatore.arcidiacono@unict.it, marco.venuti@unict.it

Lexicad: A Framework for Digital Lexicography and Early Italian Language Resources

Keywords: electronic lexicography; electronic dictionaries; Italian lexicography; diachronic lexicography; lexical resources; dictionary writing systems (DWS).

This poster describes the applications of the Lexicad lexicographic platform. Lexicad is a suite of libraries and software tools for developing electronic dictionaries and online lexical resources. Similar to frameworks used in software development, Lexicad serves as a tool to accelerate the creation of new digital lexicographic platforms, chiefly for the diachronic lexicography of early Italian.

Dictionaries Using Lexicad

The poster will present various projects that have already adopted the Lexicad system to develop their dictionary writing systems (DWS):

- **Vocabolario Dantesco (VD)** – <http://www.vocabolariodantesco.it>
- **Vocabolario Dantesco Latino (VDL)** – <http://www.vocabolariodantescolatino.it>
- **Vocabolario storico-etimologico del Veneziano (VEV)** – <http://vev.oivi.cnr.it>
- **Vocabolario del Siciliano Medievale** – <http://artesia.unict.it>
- **Dizionario Etimologico e Storico del Napoletano (DESN)** – (forthcoming publication)
- **Vocabolario del Romanesco Contemporaneo (VRC)** – (under development)

Particular emphasis will be placed on the platform, currently being published, dedicated to the **Tesoro della Lingua Italiana delle Origini (TLIO)**, the authoritative dictionary of early Italian compiled by the *Opera del Vocabolario Italiano (OVI - ovi.cnr.it)*.

This platform, named **Pluto (Piattaforma Lessicografica Unica del Tesoro delle Origini)**, goes beyond simply incorporating the over fifty thousand TLIO entries into an updated DWS; it also integrates numerous digital assets curated by the institute into a unified system.

System Architecture

The architecture of a Lexicad-based project is structured across three levels of abstraction:

1. **Core level:** This layer includes essential functions for the operation of any web platform.
2. **Linguistic and textual data processing layer:** This intermediate level contains abstract classes that describe and enable computational processing of the most common lexical entities (e.g., graphic forms, multi-word expressions, definitions, etc.).
3. **Customization layer:** The highest level in the hierarchy, dedicated to platform-specific customizations, allowing developers to add new features or extend the software libraries of the lower layers.

With the exception of the core level, Lexicad is structured into *applications* that can be easily transferred between different implementations.

The Lexicad editorial system is independent of the corpus query system (CQS) but can be easily integrated with various corpus querying tools through a **context manager**. For instance:

- Projects using the **GATTO** software developed by CNR-OVI can import linguistic contexts directly from this program.
- The **VDL** has been integrated with **Dante Search** APIs.

This means that different projects can retain their own corpus search engines while still being able to transfer data seamlessly into the Lexicad interface. Once transferred, the editorial interface allows lexicographers to query, edit, and annotate linguistic contexts to be included under dictionary entries.

Additional Applications

Thanks to its flexible architecture, Lexicad has been used not only for dictionaries but also in various *digital humanities* projects. The poster will present the following projects:

- **ItalArt** – *L'italiano delle arti tra Medioevo e Rinascimento* – <http://italart.ovi.cnr.it>
- **Rime disperse di Petrarca** – <http://rdp.ovi.cnr.it>
- **Ciclo di Guiron le Courtois** – <https://guiron.fefonlus.it>
- **Bibliografia dei Commenti Danteschi** – <http://bcd.ovi.cnr.it>

Additionally, the **LexiMap module** will be briefly described. This tool, designed for georeferencing linguistic data, was developed to manage the maps of **AGLIO – Atlante Grammaticale dell'Italiano delle Origini** (<http://aglio.ovi.cnr.it>), as part of the **MIRA - Mappatura dell'Italo-Romanzo Antico** project (<https://data.snf.ch/grants/grant/205028>), which focuses on the geolinguistic mapping of early Italo-Romance dialects

William Rolston Ashford
Dictionaries of the Scots Language, United Kingdom
william.ashford@dsl.ac.uk

Building a better Word Bank: Organising lexicographic evidence at Dictionaries of the Scots Language

Keywords: Scots language, lexicography, dictionaries, dialectology, historical lexicography

This year, Dictionaries of the Scots Language (DSL) will undertake a project to update our in-house database of citations, the DSL Word Bank (DSLWB), and adjust it to our future editorial needs.

Based on the software tool *ling.surf* (Dollinger, 2010), a web-based dictionary database and editing tool developed for *A Dictionary of Canadianisms on Historical Principles*, we built the DSLWB in 2019 to provide a reliable and easily accessible repository of evidence of Scots words and their usage.

After more than four years of working with the DSLWB 1.0 we are implementing numerous refinements and new features to ensure this core tool continues to meet our evolving needs.

This poster will provide an overview of the original DSLWB and its recent improvements.

By sharing our experience, we aim to demonstrate how bespoke software created for one dictionary project, in this case *ling.surf*, can be adapted to meet the needs of another. We also aim to provide an insight into the tools used by DSL as the editorial team look toward the future and to the prospect of adding new material to the dictionary for the first time since 2005.

References:

Dollinger, S. (2010) Software from the Bank of Canadian English as an open-source tool for the dialectologist: *ling.surf* and its features. In M. Markus, C. Upton and R. Heuberger (Eds.), *Joseph Wright's English Dialect Dictionary and Beyond: Studies in Late Modern English Dialectology* (pp. 249-261). Berne Lang.

Katarína Balleková, Peter Malčovský, Mária Mikolajová
Ľudovít Štúr Institute of Linguistics of the Slovak Academy of Sciences, Bratislava
katarina.ballekova@juls.savba.sk, peter.malcovsky@juls.savba.sk,
maria.mikolajova@juls.savba.sk

**Online popularization of research on Slovak dialects – Digital Atlas of the Slovak Language
(2024, 1.0)**

Keywords: Slovak dialects, Digital Atlas of the Slovak Language, Anton Habovštiak legacy, Digitization, Popularization

In this post, we would like to introduce the project of the digitized database of the Atlas of the Slovak Language.

The Ľudovít Štúr Institute of Linguistics of the Slovak Academy of Sciences conducted systematic research on Slovak territorial dialects across Slovakia between 1947 and 1971. The final outcome of documenting the essential features of the natural form of the national language through linguistic geography methods was the four-volume Atlas of the Slovak Language (ASL). This atlas presents phonetics, morphology, word formation, and lexicon in a thematic and cartographic form with commentary. The National Linguistic Atlas comprehensively illustrates the territorial diversity of the Slovak language. This fundamental synthetic work on Slovak dialects was published as an atlas project between 1968 and 1984 in book form, consisting of separate map sections and independent textual commentaries for each volume. However, over time, the printed edition has become almost inaccessible.

On the occasion of the 100th anniversary of Anton Habovštiak, the author of the fourth volume of ASL (1984), a new digital database representing the vocabulary of Slovak dialects was launched in 2024. This innovative solution integrates multifaceted pairing of information on cartographically recorded linguistic phenomena with descriptive explanations in the commentary section. The technical implementation of the website for the Digital Atlas of the Slovak Language drew inspiration from established international online projects. The ambition was to create a simple web platform with modern features and intuitive navigation. A key feature of the solution is its responsive web design, ensuring accessibility on smartphones, tablets, and other devices.

The database of the Digital Atlas of the Slovak Language is available to the public on the institution's website at: <https://digiASJ4.juls.savba.sk>, it is an output of the VEGA project no. 2/0114/22 Dictionary of Slovak dialects IV.

References:

- Balleková, K. (Ed.) (2024) *Digitálny Atlas slovenského jazyka / Digital Atlas of the Slovak Language* (2024, 1.0). Jazykovedný ústav Ľudovíta Štúra SAV.
<https://digiasj4.juls.savba.sk/>
- Balleková, K. (2025) Slovenské nárečia ožívajú na mapách už aj online. Ed. S. Longauerová. *Akadémia / Správy SAV*, (1), 32–35.
<https://akademia.sav.sk/slovenske-narecia-ozivaju-na-mapach-uz-aj-online/>
- Buffa, F. (1985) Štvrtý zväzok Atlasu slovenského jazyka. *Slovenská reč*, 50(5), 310–313.
<https://www.juls.savba.sk/ediela/sr/1985/5/sr1985-5-lq.pdf>

- Habovštiak, A. (1984) *Atlas slovenského jazyka. Lexika. IV/1, 2. Mapy. Úvod – komentáre – dotazníky – indexy*. Bratislava: Veda.
- Habovštiak, A. (1985) Atlasové spracovanie slovenských nárečí. *Studia Academica Slovaca*. (14), 169–189. https://zborniky.e-slovak.sk/SAS_14_1985.o.pdf

Vladimír Benko

Ľudovít Štúr Institute of Linguistics of the Slovak Academy of Sciences, Bratislava

vladimir.benko@juls.savba.sk

Needles in a Haystack? Looking for Neologisms in a Web-Crawled Corpus

Keywords: neologisms, web corpora, ensemble lemmatization

Though the role of corpora in dictionary compilation is seldom challenged, the actual procedures applied by the respective lexicographic teams can vary a lot. In general, the problem can be described as an analysis of a large amount of textual data provided by contemporary corpora, often containing several billion tokens. From the neology perspective, an additional problem appears: finding new lexical units that qualify for inclusion into the headword list most effectively.

Our paper will present a methodology applied within an ongoing project of the Orthographic—Grammatical Dictionary of the Slovak Language (Sokolová and Jarošová (Eds.), 2022) that aims to include as many new lexical items as possible. As the dictionary is basically continually being published in electronic form¹ at the Institute's Dictionary portal², the headword list can be updated as needed without too much overhead.

While several corpora are being used by our lexicographic team, the most useful one for identification of neologisms proved to be the *Araneum Slovaccum*³, a web-crawled corpus containing data since 2013 (Benko, 2014; 2024). The corpus is updated by a new crawl every 6 months, which typically brings (after filtration and deduplication) approximately 250 million of new corpus tokens.

The basic procedure of identifying neologisms is relatively straightforward: creating a lemmatized frequency list of lexical items not present in the morphological lexicon (OOVs) based on the respective dictionary and analysing the most frequent items. For large corpora, however, such lists are (firstly) typically really huge, and (secondly) contain a large proportion of items that do not qualify for becoming a dictionary headword, such as various kinds of proper nouns, abbreviations, and acronyms, frequent misspellings, etc. We will show how we cope with these issues to make the procedure feasible and efficient.

References:

- Benko, V. (2014) *Aranea: Yet another Family of (Comparable) Web Corpora*. In *Text, Speech and Dialogue: 17th International Conference (TSD 2014)*, Brno, Czech Republic, September 8-12, 2014. Proceedings. - Springer, 2014, pp. 247-256. ISBN 978-3-319-10815-5.
- 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. ISBN 978-3-031-70562-5. ISSN 0302-9743.
- Sokolová, M. and Jarošová, A. (Eds.) (2022) *Ortograficko-gramatický slovník slovenčiny. A – Ž (používateľská verzia Slovníka súčasného slovenského jazyka)*. (Lexicographic team: D.

¹ A printed version of the of the dictionary, however, is planned to be published by the end of this year.

² <https://slovník.juls.savba.sk/?w=lexikografia&s=exact&c=U062&cs=&d=ogs#>

³ <http://unesco.uniba.sk/guest/>

Guričanová, M. Chochol, M. Kopecká, T. Laliková, S. Mislovičová, R. Ondřejková, I. Vančová).

Attila Benő¹, Katalin P. Márkus², Tibor M. Pintér³

¹Babeş–Bolyai University, Cluj-Napoca, Romania

Department of Hungarian and General Linguistics

²Károli Gáspár University of the Reformed Church in Hungary, Budapest

Department of English Linguistics

³Károli Gáspár University of the Reformed Church in Hungary, Budapest

Department of Hungarian Linguistics

attilabe@yahoo.com, p.markus.kata@kre.hu, m.pinter.tibor@kre.hu

A Hungarian Dictionary in the Online Era

Termini Online Hungarian–Hungarian Dictionary and Database

Keywords: dictionary editing, practical lexicography, online dictionaries, Hungarian language, spoken language

The Termini Online Hungarian–Hungarian Dictionary and Database describes the lexicon of the Hungarian language as spoken in the countries surrounding Hungary. It is considered to be a general dictionary of Present-day Hungarian and stands as a resource for several linguistic research conducted on contact varieties of Hungarian. Each entry contains authentic sample sentences to illustrate the use of the headword, making it possible to examine the special use of a word or construction in a grammatical and pragmatic context. The lexicographical database is edited online in eight countries. The presentation is about to outline some best practices of the editing process. Online editing makes it possible for the dictionary to expand – even simultaneously – as a result of activity in eight countries. In the presentation the novelties and peculiarities of the dictionary will be highlighted, touching on the following topics: dictionary structure, IT support, database character, multimedia elements, and labelling system.

References

- Benő, Attila, Juhász, Tihamér & Lanstyák, István (2020) A Termini „határtalan” szótára [The “Borderless” Dictionary of the Termini Research Network]. *Magyar Tudomány*, 2, 153–163.
- Lanstyák, István, Benő, Attila & Juhász, Tihamér (2010) A Termini magyar–magyar szótár és adatbázis [The Termini Hungarian–Hungarian dictionary and database]. *Regio*, 21(3), 37–58.
- M. Pintér, Tibor, P. Márkus, Katalin & Benő, Attila (2023) A Dictionary for Hungarian Varieties Spoken in the Carpathian Basin. *Acta Universitatis Sapientiae Philologica*, 15(2), 166–181.
- M. Pintér, Tibor (2019a) Magyar nyelv a határon. A Termini magyar–magyar szótár. 1. rész. [Hungarian language on the border. The Termini Hungarian–Hungarian dictionary 1.]. *Magyar Nyelv* 115 (3) 333–347. <https://doi.org/10.18349/MagyarNyelv.2019.3.333>
- M. Pintér, Tibor (2019b) Magyar nyelv a határon. A Termini magyar–magyar szótár. 2. rész. [Hungarian language on the border. The Termini Hungarian–Hungarian dictionary 2.]. *Magyar Nyelv* 115 (4) 473–479. <https://doi.org/10.18349/MagyarNyelv.2019.4.473>

Wiebke Blanck
Friedrich-Alexander-Universität Erlangen-Nürnberg, FAU, Germany
wiebke.blanck@fau.de

Towards the pragmatic item *swearword* and its (current and future) role in general-language online dictionaries of German

Keywords: swearwords, hate speech, online dictionaries, lexicography, pragmatic item, pragmatic value

Swearwords form a separate class of words (cf. Schippan 2002: 88), usually denoting pejorative or insulting statements about people (WLWF III: 703). Typical lexical areas are excrement, pathology and sexuality. They are often given the pragmatic item ‘swearword’ in (online) dictionaries, although the selection criteria are not always comprehensible (ibid.: 704). The representation of swearwords in general-language online dictionaries of German varies considerably. A review of the examples from Seibicke (1991: 1190) – *Affe* and *Arschficker* – shows this by the example of DWDS: In the case of *Affe*, the pejorative meaning is marked by the pragmatic value ‘abusive word’, whereas *Arschficker* is labelled as ‘vulgar, pejorative’. Thus, several labels are used for the same pragmatic value. Duden online shows similar results. Regardless of the conception of dictionaries and variations that may occur over long project periods, the question arises whether the labels have been set rather arbitrarily. Considering Schippan (2002), it would be possible to label both words as swearwords or to only indicate descriptive pragmatic items of the type *vulgar, pejorative*, etc. These single examples illustrate a tendency of general-language online dictionaries to use inconsistent labels. This contribution, firstly, questions whether the pragmatic item *swearword* is still appropriate for documenting pejorative language in online dictionaries, given the inconsistency in the use of this label. Secondly, it points to lexicological concepts that do not situate pejorative language exclusively at the individual word level and should be considered in (meta-)lexicography. Guillén-Nieto (2023) takes this approach regarding *hate speech*; in reference to Hom (2008) she shows the central role of epithets in the construction of pejorative language. Likewise, Marx/Meier-Vieracker (2024: 440) point to the complexity of utterances that fall within the scope of *hate speech*. Against this background, this contribution suggests focussing more on contexts when displaying pejorative language in online lexicography: avoid the label *swearword* and display pragmatic complexity by integrating a section in the article that shows several (recent) examples of a pejorative uses of the word (i.e. corpus findings including several text types).

References:

- Guillén-Nieto, V. (2023) *Hate Speech: Linguistic Perspectives*. de Gruyter.
Hom, C. (2008) The Semantics of Racial Epithets. *Journal of Philosophy* 105(8): 416–440.
Marx, K., Meier-Vieracker, S. (2024). Digitale Gewalt: Formen und interaktive Verfahren. In J. Androutsopoulos, F. Vogel (Eds.), *Handbuch Sprache und digitale Kommunikation* (pp. 435–454), de Gruyter.
Schippan, T. (2002) *Lexikologie der deutschen Gegenwartssprache*. Max Niemeyer.
WLWF = H. E. Wiegand, R. H. Gouws, M. Kammerer, M. Mann, W. Wolski (2020) *Wörterbuch zur Lexikographie und Wörterbuchforschung* (3), I – U. de Gruyter.

Edmond Cane
Beijing International Studies University, China
ecane2000@yahoo.com

A Construction-based model for language processing

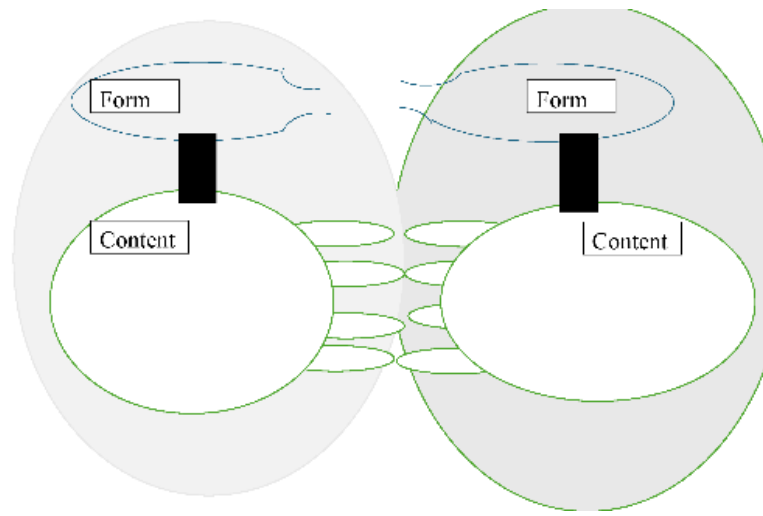
Keywords: construction grammar, constructional pairing, corpus annotation, LLM

Current annotation models in linguistic corpora and large language models (LLMs) have largely been developed within structuralist frameworks, relying primarily on tagsets. As Croft (2024) points out, many non-constructional theories of syntax depend heavily on syntactic notation systems, which ultimately limit their explanatory power. LLM accuracy and efficiency hinge on how algorithms interpret grammatical information in texts through tags and how these tags are interconnected to form a coherent system. These systems rely on co-occurrence statistics drawn from massive datasets of prior usage. This approach, however, differs fundamentally from how human speakers process language. Crucially, such models lack grammatical understanding of either input or output and therefore fail to evoke the experiential dimension of language for speakers, interlocutors, or third-party others.

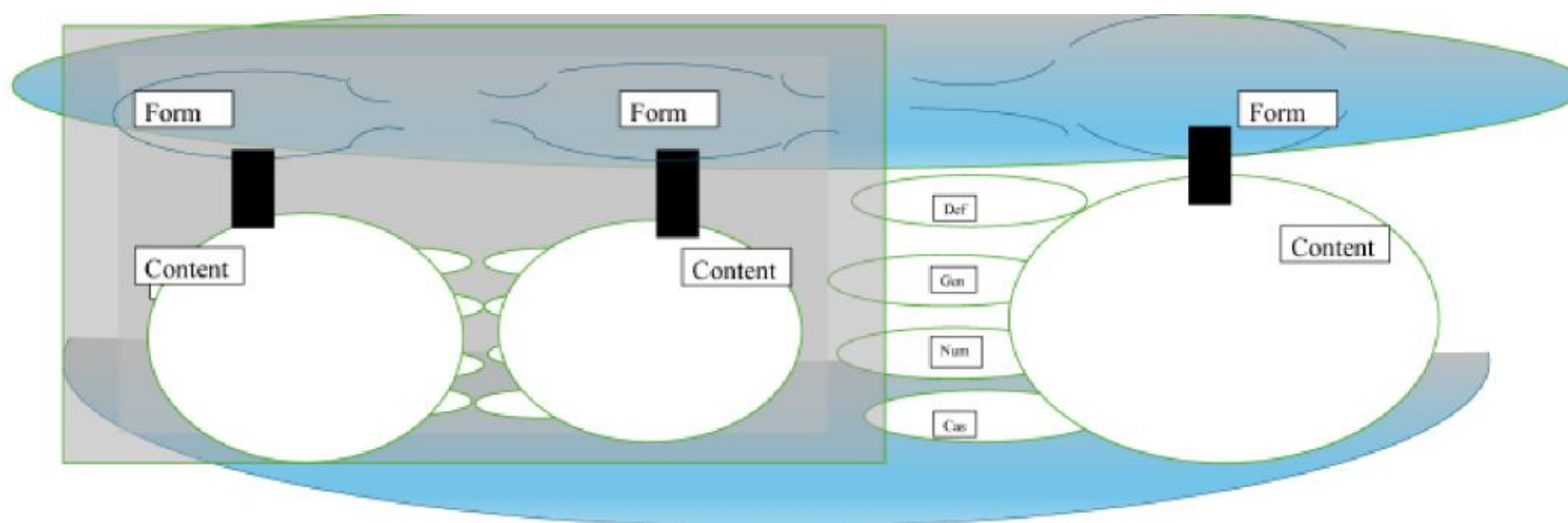
While some scholars (e.g., Manova 2023) advocate for linguistics to embrace ChatGPT-like perspectives on language processing, my research aligns with constructionist approaches (Goldberg 1995, 2006; Croft & Cruse 2004; Diessel 2019). Here, constructions—defined as form–content pairings—constitute the fundamental building blocks of language. Even morphs are treated as independent constructions, diverging from certain construction-based frameworks such as Construction Morphology (Booij 2010), Fluid Construction Grammar (Steels 2011), and Croft & Cruse’s (2004) definition, but aligning with Goldberg (2006).

In this model, all levels of constructions are represented as dictionary entries in a key-value format. The value is instantiated as a `compound_set`, which consists of i) `pos_set` – contains part-of-speech (POS) information, guiding the processing along predicted pathways, and ii) `feature_set` – for nouns and noun-related categories, this is structured as a set of four dictionaries: Number, Gender, (In)definiteness, Case. These features are not independent components but rather designed as an integrated set, ensuring cohesion and interdependence in morphosyntactic representation.

Following the constructionist principle that “it is constructions all the way down” (Goldberg 2006), all relevant information is contained within each construction, allowing processing without recourse to rule-based algorithms. Features are designed in slot-shaped format: stems and markers are themselves constructions, and when their features are compatible, they unify to yield a higher-level complex construction. See Fig. 1



Complex constructions in turn open toward unification with other constructions. If the features and internal information across constructions are compatible, they are licensed, and unification proceeds upward. (See fig. 2 how unification proceeds)



Thus, the key challenge lies in **information packaging**: designing constructions so that they can activate and integrate one another in ways that successfully process linguistic text, thereby enabling genuine “reading” of language rather than surface-level tagging.

References:

- Booij, Gert (2010) Construction morphology. *Language and linguistics compass*, 4 (7) 543-555.
- Croft, William., & Cruse, D. Alan. (2004). *Cognitive linguistics*. Cambridge University Press.
- Diessel, Holger. (2019) *The grammar network*. Cambridge University Press.
- Goldberg, A. E. (1995) *Constructions: A construction grammar approach to argument structure*. University of Chicago Press.
- Goldberg, Adele. E. (2006) *Constructions at Work: The Nature of Generalizations in Language*. Oxford University Press.
- Haspelmath, Martin. (2006) Against markedness (and what to replace it with). *Journal of linguistics*, 42 (1) 25-70.
- Langacker, Ronald. W. (1987) Foundations of cognitive grammar: Volume II: Descriptive application. Steels, Luc. (2011). Design patterns in fluid construction grammar. *Design Patterns in Fluid Construction Grammar* 1-344.

István Csernicskó¹, Krisztián Váradi²

¹ Ferenc Rákóczi II Transcarpathian Hungarian College of Higher Education, Ukraine;
University of Pannonia, Hungary

² Ferenc Rákóczi II Transcarpathian Hungarian College of Higher Education, Ukraine;
University of Pannonia, Hungary

csernicsko.istvan@kmf.org, uavaradi.krisztian@kmf.org.ua

The Role of Regional Dialects in the Development of an Educational Terminology Database and the Termini Hungarian–Hungarian Dictionary

Keywords: educational terminology, Hungarian minority, loanwords, regional dialects, translation

As part of the Science for the Hungarian Language National Programme of the Hungarian Academy of Sciences, specifically within the Hungarian Terminology Strategy Subprogramme, the Educational Terminology Subproject was launched to develop a comprehensive educational terminology database. The aim of this initiative is to standardise educational terminology across all levels of the education system in Hungary and in Hungarian-speaking regions of the Carpathian Basin. The rationale for creating the database lies in the fact that, following the Treaty of Trianon in 1920, Hungarian minorities in the annexed territories have been using numerous expressions unfamiliar either in Hungary or in the neighbouring states. The construction of the database follows a three-level approach: first, Hungarian educational terms are translated into the official languages of Hungary's neighbouring countries; second, country-specific terms are translated into Hungarian; and third, educational dialectal expressions and loanwords used in cross-border regions are documented. In the first year of development, the primary focus was on terminology related to public education. During this phase, we matched 849 Hungarian public education terms with their Ukrainian equivalents and also developed numerous candidate terms. Subsequently, we collected and translated 100 Ukraine-specific educational terms into Hungarian and documented 70 region-specific loanwords from the Transcarpathian variety of the Hungarian language. In our presentation, we aim to summarise the results of this initial phase of the database's development from the perspective of the Transcarpathian region, highlighting key semantic discrepancies between the public education systems of Hungary and Ukraine. Furthermore, the study examines the contribution of the Educational Terminology Subproject to the enrichment of the Termini Hungarian–Hungarian Dictionary.

Sara de Piniés de la Cuesta, Javier Serrano Peinado

Universidad Complutense de Madrid, Spain

Saradepi@ucm.es, Javise02@ucm.es

Identifying Collocations in Oceanographic Lexicology: A Focus on Temperature, Pressure, and Salinity

Keywords: collocations, domain-specific, Lexical Priming, colligations, oceanography

This study explores the linguistic patterns of the oceanographic discipline by applying the Lexical Priming Theory (Hoey, 2005), focusing on the collocations of temperature, pressure, and salinity. The term collocation refers to the pairing of at least two words, which “appear frequently in each other’s company” (Hoey, 2005, p. 2), pointing to scientific discourse and conceptual precision. Besides, the Lexical Priming theory postulates that “repeated encounters in particular contexts significantly influence word associations” (Hoey, 2005, p. 45), linking it with domain-specific patterns.

The collocation extraction procedure employed a corpus-based methodology, which comprised building a corpus of oceanographic research articles selected from top ranked oceanographic journals, and the categorization of each document regarding its scientific discipline (biology, chemistry, geology, and physics). The corpus was explored with Sketch Engine, a corpus analysis tool that listed the collocations of the given words (Kilgariff et al., n.d.). As a result, collocation structures like “adjective + noun”, “noun + noun”, or “preposition + noun” were found to be the most preferred ones, showing also the distribution of the collocations across the different scientific disciplines.

Having filtered the collocates using statistical association measures (Xiao, 2015), it was found that “temperature + salinity” is a very strong and reciprocal collocate. Moreover, “pressure” was found to have a very strong collocate with the words “air”, “fishing” or “predation”. Other findings showed that “modifier+ lemma” is the most common structure for the three given terms: “pressure” (60%), “temperature” (48,55%) and “salinity”(37,67%). The collocational distribution across disciplines points out their function as key concepts in the oceanographic domain, because the term “salinity” appears widely and significantly distributed across disciplines, “temperature” outperforms in “Biology”, whereas “pressure” is equally distributed between “physics” and “biology”.

These results show the existence of domain-specific lexical patterns that distinctively differentiate each term.

References:

- Hoey, M. (2005) *Lexical Priming: A New Theory of Words and Language*. Taylor & Francis Group. <http://ebookcentral.proquest.com/lib/universidadcomplutense-ebooks/detail.action?docID=178100>
- Kilgariff, A., Rychly, P., Smrz, P., & Tugwell, D. (n.d.). The Sketch Engine.
- Xiao, R. (2015) Collocation. In D. Biber & R. Reppen (Eds.), *The Cambridge Handbook of English Corpus Linguistics* (pp. 106–124). Cambridge University Press. <https://doi.org/10.1017/CBO9781139764377.007>

Csilla Ilona Dér
ELTE Research Centre for Linguistics, Hungary
der.csilla@nytud.elte.hu

Interjection-discourse markers in 19th–20th century Hungarian dictionaries, grammars and databases

Keywords: interjection, discourse marker, interactives, categorization, emotive/volitional functions

This paper examines the relationship between primary interjections and discourse markers in Hungarian. Primary interjections are those that are monosyllabic and conceptually non-referential (Ameka 1992), e. g. *ó* 'oh', *jaj* 'ouch'. These elements have been studied from various perspectives in the international literature (cf. Ameka 1992, Heine 2023: 181–199). Some approaches prioritize part-of-speech classifications, while others interpret them as exclamatory utterances or unstructured sentences (cf. Kugler 2000, Dingemanse 2023). In line with Heine's (2003: 7) functional conceptualization, interjections are categorized as semantically, syntactically, and prosodically independent, invariant deictic/indexical forms that are anchored in the discourse situation. Among these, interjections are the elements that "express a sudden change in the speaker's emotional or cognitive state" (Heine 2003: 185). Heine (2023: 127–140, 280) also considers discourse markers to be interactives, with interjection-derived discourse markers being viewed as interactives that have undergone grammaticalization. Other scholars discuss the use of interjections as discourse markers, treating the former as a part of speech and the latter as a functional category (e.g., Montes 1999, Norrick 2009, Roggia 2012). Indeed, interjections often do not stand alone but accompany other structured or unstructured sentences, and they can play various interpersonal, sequential, etc. roles in discourse organization (cf. Crible 2018). This study focuses on interjections that have undergone grammaticalization into discourse markers, specifically *ó~oh*, *jaj~ajaj*, and *á~áh~ah*. We analyze 19th- and 20th-century printed and online Hungarian dictionaries (CzF, Értész, ÉKsz, Nszt), as well as grammars (e.g., Versegly 1818, Fogarasi 1843) and databases (MTSz, MNSz2, BEA), in terms of how these elements are categorized and whether they exhibit discourse marker characteristics in both their descriptions and actual language use. Additionally, we clarify how the emotive and volitional functions associated with interjections correspond to discourse marker features and how these two categories should be represented in dictionary descriptions.

References:

- Ameka, F. K. (1992) Interjections: The universal yet neglected part of speech. *Journal of Pragmatics*, 18(2), 101–118. [https://doi.org/10.1016/0378-2166\(92\)90048-G](https://doi.org/10.1016/0378-2166(92)90048-G)
- BEA = *Beszélt Nyelvi Adatbázis [Hungarian Spontaneous Speech Database]*. (n.d.). MTA Nyelvtudományi Intézet <https://fonetika.nytud.hu/bea-beszelt-nyelvi-adatbazis/?lang=en>
- Crible, L. (2018) *Discourse markers and (dis)fluency: Forms and functions across languages and registers*. Amsterdam–Philadelphia: John Benjamins.
- CzF. = Czuczor, G., & Fogarasi, J. (Eds.) (1862) *A magyar nyelv szótára [Dictionary of the*

- Hungarian Language*]. Pest: Emich Gusztáv Magyar Akadémiai nyomdásznál
<https://www.arcanum.com/hu/online-kiadvanyok/Lexikonok-a-magyar-nyelv-szotara-czuczorfogarasi-55BEC/>
- Dingemanse, M. (2023) Interjections. In E. van Lier (Ed.), *The Oxford handbook of word classes* (pp. 477–492). Oxford: Oxford University Press.
- ÉrtSz = MTA Nyelvtudományi Intézet. (Ed.) (1961) *A magyar nyelv értelmező szótára* [*The Explanatory Dictionary of the Hungarian Language*]. Budapest: Akadémiai Kiadó
<https://www.arcanum.com/hu/online-kiadvanyok/Lexikonok-a-magyar-nyelv-ertelmezo-szotara-1BE8B/>
- ÉKsz = Pusztai, F. (Ed.) (2003) *Magyar értelmező kéziszótár* [*Hungarian Explanatory Handbook Dictionary*]. Budapest: Akadémiai Kiadó.
- Fogarasi, J. (1843) *A magyar nyelv szelleme I. kötet: Művelt magyar nyelvtan elemi része* [*The spirit of the Hungarian language Vol. 1: Elementary part of educated Hungarian grammar*]. Pest: Heckenast Gusztáv.
- Heine, B. (2023) *The grammar of interactives*. Oxford: Oxford University Press.
- Kugler, N. (2000) Az indulatszó [Interjections]. In B. Keszler (Ed.), *Magyar grammatika* (pp. 295–296). Budapest: Nemzeti Tankönyvkiadó.
- MNSZ = *Magyar Nemzeti Szövegtár* [*Hungarian National Corpus*]. (n.d.)
<http://clara.nytud.hu/mnsz2-dev/>
- MTSZ = *Magyar Történeti Szövegtár* [*Hungarian Historical Corpus, 1772–2010*]. (n.d.)
<http://clara.nytud.hu/mtsz/>
- Montes, R. G. (1999) The development of discourse markers in Spanish: Interjections. *Journal of Pragmatics*, 31(10), 1289–1319. [https://doi.org/10.1016/S0378-2166\(98\)00107-1](https://doi.org/10.1016/S0378-2166(98)00107-1)
- Nszt. = Ittész, N. (Ed.) (2006) *A magyar nyelv nagyszótára* [*The Great Dictionary of the Hungarian Language*]. Budapest: MTA Nyelvtudományi Intézet.
<https://nagyszotar.nytud.hu/index.html>
- Norrick, N. R. (2009) Interjections as pragmatic markers. *Journal of Pragmatics*, 41(5), 866–891. <https://doi.org/10.1016/j.pragma.2008.08.005>
- Roggia, A. B. (2012) Eh as a polyfunctional discourse marker in Dominican Spanish. *Journal of Pragmatics*, 44(13), 1783–1798. <https://doi.org/10.1016/j.pragma.2012.08.012>
- Schiffrin, D. (1987) *Discourse markers*. Cambridge: Cambridge University Press.
- Versegly, F. (1818) *Magyar grammatika, avagy nyelvtudomány* [*Hungarian grammar, or linguistics*]. Buda: n. p.

Fazakas Emese¹, Zsemlyei Borbála²

¹Department of Hungarian and General Linguistics, Babes–Bolyai University, Cluj-Napoca, Romania

²Department of Hungarian and General Linguistics, Babes–Bolyai University, Cluj-Napoca, Romania

emese.fazakas@ubbcluj.ro, borbala.zsemlyei@ubbcluj.ro

The necessity of an index volume to the *Historical Dictionary of the Hungarian Language in Transylvania*

Keywords: the Historical Dictionary of the Hungarian Language in Transylvania, index volume, entry words, editorial principles, editorial process

The concept of editing a special dictionary which would cover the complete vocabulary of the Hungarian language used in Transylvania is closely linked to the linguist Attila Szabó T. (1906–1987). In the 1920s, Attila Szabó T. – as a student of the archivist Lajos Kelemen – started gathering data in the archives throughout Transylvania for a dictionary that would present the old legal terms used in all kinds of legal processes in Transylvania. The more documents he analysed, the more convinced he became that it would be a shame not to do anything with the language material contained by all those documents. The result of a century of hard work is the *Historical Dictionary of the Hungarian Language in Transylvania*, which gives a complete account of the Hungarian language of the period between the 15th and the 19th century of Transylvania in 14 volumes.

However, due to the large amount of language data, the large number of editors, to the lack of technology, the historical and political changes, etc., there are many inconsistencies and inadequacies in the dictionary, and sometimes it is difficult for readers, researchers to use it. For example, even the concept of the entry word changed throughout the 14 volumes. At first Attila Szabó T. included suffixated words (eg.: *egyemben*, *fennszóval*, *másszor* etc.), phrases (eg.: *herbatea-főző ibrik*, *hordószorító vasabroncs* etc.) etc. as entry words which in the later volumes were omitted. Besides there are many dialectal and/or historical variants, individual older and modern dialectal forms as entry words that is difficult for researchers to find since they will primarily remember the today's standard version.

Thus, it became obvious that after the completion of this monumental work, an index volume – which would serve as a map to the huge amount of language data of the 14 volumes – is of outmost importance. As the editors of this volume, we aim to present on the one hand the editorial principles, and on the other, the whole editorial process of the index volume.

Márta Fischer¹, Dóra Mária Tamás²

¹Budapest University of Technology and Economics, Budapest, Hungary

²ELTE Research Centre for Linguistics, Research Institute for Lexicology, Budapest, Hungary
fischer.marta@gtk.bme.hu, tamas.dora.maria@nytud.elte.hu

Educational Terminology in Hungary and the Hungarian-speaking Communities of the Neighbouring Countries: Strategic Issues, Dilemmas and Solutions

Keywords: educational terminology, Hungarian-speaking communities, terminology strategy, termbase, regional equivalents, lingua academica

The presentation will provide an insight into the peculiar terminological situation of Hungary and the seven Hungarian-speaking communities in the Carpathian Basin, and the efforts made in this regard. It will be argued that the phenomenon of linguistic fragmentation underscores the need for the systematization of educational terminology and its publication in a terminological database. The development of educational terminology is instrumental in achieving several objectives of language policy and terminology strategy, including the reinforcement of the right of minorities to use their mother tongue, the facilitation of educational mobility between the mother country and the regions, and the provision of support to local minority language communities through the recording and publication of regional equivalents, in accordance with the Language Management Theory. In view of its role as a lingua franca and a lingua academica, the creation of a terminological database will also enable to incorporate English, thus reflecting the principle of additive bilingualism. In order to ensure the inclusion of useful data, the staff and experts of the HUN-REN Hungarian Research Centre for Linguistics have contacted the Hungarian Educational Authority, within the framework of the Hungarian Terminology Strategy sub-programme, to define the scope of relevant data. The significance of defining the strategic objectives of the project is reflected at the micro-level of the database, in the careful selection of data categories and contents. The quality of the compilation and the editing of the database requires the adherence to terminological principles and methods. The development of an editing manual and the teaching of terminological principles to all involved is therefore essential.

Emmanuel Sylvain Fomat
Department of English and German Philology
University of Santiago de Compostela, Spain
sylvainfomat4@gmail.com

**A lexico-semantic analysis of neologisms in the Anglophone crisis in
Cameroon**

Keywords: neologisms, Anglophone crisis, identity, resistance, sociopolitical discourse

The ongoing Anglophone crisis in Cameroon, which began in late 2016 with protests by Anglophone lawyers and teachers against perceived marginalization by the Francophone-dominated state, has led to the development of new terms reflecting the complex sociopolitical and cultural dynamics of the conflict. By analyzing neologisms such as “Ambazonia”, “Ghost Town”, “Odeshi”, or “Ground Zero”, this study highlights their role in shaping identity, expressing resistance, and capturing the essence of the crisis. It seeks to answer the following question: What are the various categories of neologisms that have emerged during the Anglophone crisis, and how do these new lexical items shape discourse, identity, and public understanding of the conflict? The research uses both lexico-semantic analysis and Critical Discourse Analysis (CDA) to investigate the formation and meanings of these terms, focusing on the ways in which they convey political ideologies, social realities, and cultural expressions. It analyzes neologisms from news articles, social media, interviews, and academic literature to understand how language navigates and frames the crisis. The findings show that neologisms are not only linguistic innovations but also powerful tools for articulating the struggle for autonomy, shaping public discourse, influencing group identity, and framing the narrative of resistance. This study contributes to lexicography, sociolinguistics, and conflict communication, offering insights into how language evolves in response to political crises. It also emphasizes the need for further documentation of neologisms in crisis situations and their broader implications for identity construction, sociopolitical discourse, and conflict resolution.

References:

- Baker, P., & McEnery, T. (2015) *Corpora and discourse studies: Integrating discourse and corpora*. Palgrave Macmillan. <https://doi.org/10.1057/9781137431738>
- Chilton, P., & Schäffner, C. (2011) Discourse and politics. In T. A. van Dijk (Ed.), *Discourse studies: A multidisciplinary introduction* (2nd ed., pp. 303–330). SAGE Publications Ltd. <https://doi.org/10.4135/9781446289068.n15>
- Dijk, T. A. van. (2008) *Discourse and power*. Palgrave Macmillan. <https://doi.org/10.1007/978-1-137-07299-3>
- Fairclough, N. (1995) *Critical discourse analysis: The critical study of language*. Longman. <https://doi.org/10.4324/9781315834368>
- Nkwain, J. (2022) Current insights into the evolution of Cameroon English: The contribution of the ‘Anglophone problem’. *Athens Journal of Humanities & Arts*, 9(3), 233–260. <https://doi.org/10.30958/ajha.9-3-3>

Péter Gaál

University of Maribor, Faculty of Arts, Department of Hungarian Language and Literature,
Slovenia

gaal.peter.hun@gmail.com

Dictionary use habits of teachers from the Prekmurje region - A case study from Slovenia

Keywords: Hungarian ethnic minority in Slovenia, dictionary use, dictionary preferences, lexicographical needs, didactics of dictionary use

The paper summarises the experiences of a study into dictionary use based on an online questionnaire survey of teachers in Hungarian-Slovenian bilingual schools in Slovenia. The Hungarian minority in the Prekmurje region of Slovenia has had access to bilingual education since the late 1950s: three primary and one secondary school are available for those who wish to continue their studies in Hungarian in Slovenia. Bilingual education is both an opportunity and a challenge, providing a chance for the Hungarian-speaking community in Slovenia to survive, or at least to slow down the decline of the community. In the bilingual field, monolingual and bilingual dictionaries can be a useful tool for dealing with various language problems, and it is primarily the teachers in bilingual schools who have to familiarise students with them. However, no data are available on which dictionaries are used by teachers, what their habits of use are, or how they might use them in their work. There is some research on dictionary use in Hungary (e.g. Gaál 2017, 2020 and P. Márkus 2024) for students and teachers in public education, but no similar research has been conducted in Hungarian communities beyond the borders. The aim of the research presented in this paper is to explore the dictionary use habits of teachers in bilingual primary and secondary schools in the Prekmurje region. The results will provide data on the types of dictionaries used by teachers, the most common dictionary use situations and purposes, the role of dictionaries in education, their current issues of dictionary use and lexicographic needs.

References:

- Gaál, P. (2017) Online-szótár-használat a szombathelyi ELTE Bolyai János Gyakorló Általános Iskola és Gimnáziumban: Négy interjú tapasztalatai In Hajba, R.; Tóth, P. (Eds.) *A véges végtelen: Tanulmányok Vörös Ferenc 60. születésnapjára*. Savaria University Press pp. 71 –78.
- Gaál, P. (2020) Középiskolás tanulók szótárhasználati szokásai – egy vas megyei kérdőíves felmérés eredményei. *Alkalmazott Nyelvtudomány*, 20 (2) 1–19.
https://alkalmazottnyelvtudomany.hu/wordpress/wp-content/uploads/Gaal_tan.pdf P.
- Márkus, K. (2024) *Szótárdidaktika az idegennyelv-oktatásban*. Tinta Könyvkiadó.

Radovan Garabík, Agáta Karčová

Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences, Bratislava, Slovakia
radovan.garabik@kassiopeia.juls.savba.sk, agata.karcova@korpus.juls.savba.sk

Analyzing grammatical anomalies in lexical data for fun and profit

Keywords: corpus, grammar, statistical anomalies, corpus-driven, dictionary

The Dictionary of Adjectives in Slovak (Prídavné mená v slovenčine – štúdie a štatistiky na báze korpusov slovenčiny; in print) is a corpus-driven comprehensive dictionary of Slovak adjectives (https://www.juls.savba.sk/pub_adjektiva_2025.html). The dictionary includes a new analysis of declension patterns, complete paradigms of prototypical adjectives, and an analysis of anomalous occurrences of the adjectives in the corpus.

By an “anomalous occurrence” we mean a deviation from a typical or expected statistical distribution of the set of values of a given grammatical category. Such an anomaly is often present in (and caused by) fixed phrases and idioms. For example, a given word (noun or adjective) may occur more frequently or exclusively in a specific grammatical case compared to other words of the same part of speech (POS). The corpus from which we extract the grammatical categories is automatically lemmatized and tagged with a morphosyntactic description (MSD) (Garabík & Mitana 2023).

We model the distribution of grammatical categories by assigning each category a vector g_w for a word w in an n -dimensional space, where the coordinate axes correspond to the values of the grammatical category, and the coordinates to the number of occurrences of a word in a given grammatical category value (e.g. a category can be a case, a category value is nominative, genitive etc.). The dimension n corresponds to the number of values within that category (e.g., the number of cases, degrees, genders, etc.). We assume a hypergeometric probability distribution of the words in corpus and use the Clopper-Pearson confidence interval to filter out less certain values. We calculate cosine similarity between the vector g_w and a vector G averaged over all the words of the same POS, and use the similarity to quantify the most anomalous occurrences. After normalizing the vectors, we can use simple coordinate difference to rank the anomalies by their “strength”. In the dictionary we analyze anomalies in case, number, gender, degree and several synthetic morphologically significant combinations. The results are presented in an easy to read, intuitive lexicographical format.

The method as described is sufficiently generic and can be used for any POS in any medium to highly inflected language, as long as a sufficiently large lemmatized corpus, tagged for POS and MSD is available. We also automatically extract short meaningful examples of the anomalies from the corpus, using a Monte Carlo-like sampling method, ranking examples by their frequency and desired length. The extraction process is easily generalized to obtain examples of a given lemma, word form, word with a certain value of a grammatical category, or an arbitrary CQL query supported by the target corpus.

We implemented a simple public web interface for the above-mentioned example extraction, querying the ARANEA family of web corpora (Benko 2024), we currently support ARANEA corpora for the following languages: Bulgarian, Czech, Dutch, English, Estonian, Finnish, French, Georgian, Hungarian, Italian, Latin, Latvian, Persian, Polish, Romanian, Russian, Slovak, Spanish, Swedish, Ukrainian, and Uzbek, as well as the representative Slovak National Corpus, version prim-11. The webpage is available at the address

<https://www.juls.savba.sk/typicalities/>

References:

- Benko, V. (2024) The Aranea Corpora Family: Ten+ Years of Processing Web-Crawled Data. In E. Nöth, A. Horák, & P. Sojka (Eds.), *Lecture Notes in Computer Science: Text, Speech, and Dialogue*, Vol. 15048, (pp. 55–70). Springer.
- Garabík, R., & Mitana, D. (2023) Analysing accuracy of Slovak language lemmatization and MSD tagging. *Slovenská reč*, 88(2), 129–140.
- Karčová, A., & Garabík, R. (2025) *Prídavné mená v slovenčine: Štúdie a štatistiky na báze korpusov slovenčiny*. [In print]

Acknowledgements:

This work received support from the CA21167 COST action UniDive, funded by COST (European Cooperation in Science and Technology).

The dictionary received support from the project Skloňovanie prídavných mien v slovenčine s korpusovými príkladmi, funded by the Slovak Academy of Sciences Return Project Scheme for Parents Returning to Work after Maternity or/and Parental Leave.

This work is supported by the project VEGA No. 2/107/23 Word Embeddings of Vague Expressions in Slovak.

Luisa Giacomà
Università della Valle d'Aosta, Italy
l.giacoma@univda.it

The strange case of the lexicographic paratext

Keywords: bilingual dictionaries, monolingual dictionaries, paratext, German, Italian

The bilingual German-Italian dictionary consists of three parts: the German section, the Italian section and the paratext. The latter, however, is mysteriously missing in dictionaries that can be consulted online. The present contribution therefore intends to investigate first of all the causes of this phenomenon, which has its roots in a publishing tradition that is generally disinclined to value the paratext.

Furthermore, looking at the text accompanying a dictionary, one notices the extreme heterogeneity of this textual typology. If one analyses, for example, the monolingual dictionary *Deutsche Idiomatik*. (Schemann 1993), as linguists one cannot but rejoice at the almost 100-page introduction in which the author describes the work in great detail and situates it within the scientific debate. The work forms the basis for several bilingual editions, including the *German-Italian idiomatic dictionary* (2009). In the Italian edition, one observes a drastic cut of the introduction, reduced to only four pages, conceived as mere 'instructions for use'. Certainly, the Italian publisher's choice was influenced by the knowledge of its audience, which notoriously does not read dictionary introductions. The (lack of) didactics of dictionaries, repeatedly emphasised by linguists, also influences this unsatisfactory aspect of the relationship between the public and dictionaries.

Subsequently, through a comparative analysis of the paratexts of the most widely used German-Italian dictionaries of more or less recent times, I intend to draw an initial state of the art, in the hope that this will induce editors and the public to get to know and appreciate them more.

Anna Grzeszak
University of Warsaw, Poland
anna.grzeszak@uw.edu.pl

On unconventional multilingual dictionaries (17th–21st century)

Keywords: polyglot dictionaries, metalexicography, dictionary typology, equivalents, dictionary structure

Multilingual dictionaries are typically defined as dictionaries that either 1) cover more than two languages (e.g. Fuertes-Olivera & Bergenholtz, 2020; Haensch, 1991) or 2) coordinate equivalent lexical units of more than two languages (e.g. Hartmann & James, 1998; Zgusta, 1971). These two definitions are often treated synonymously in the metalexicographic literature. However, the first definition is broader in scope than the second. This study focuses on dictionaries that can be classified as multilingual solely based on the first definition. In other words, it concerns dictionaries that embrace at least three languages but do not present equivalents for lexical units of all of these tongues. The aim of the presentation is to answer the following questions: 1) what are some examples of such unconventional multilingual dictionaries, and 2) how is multilingualism realized in them.

The first step in this research involved compiling a list of ten dictionaries that meet the specified criteria, followed by an analysis of their contents. Both contemporary and early dictionaries were considered. The list was compiled based on multilingual lexicography literature (i.a. Domínguez Vázquez et al., 2020; Haensch, 1991) and supplemented with findings from the author's own research. In the ten identified dictionaries, multilingualism is realized in two distinct ways. Some are multi-part dictionaries, which are collections of several bilingual or monolingual dictionaries. Others are dictionaries with a simple structure, in which at least three different languages alternately serve as the source language.

References:

- Domínguez Vázquez, M. J., Mirazo Balsa, M., & Valcárcel Riveiro, C. (Eds.). (2020) *Studies on multilingual lexicography*. de Gruyter.
- Fuertes-Olivera, P. A., & Bergenholtz, H. (2020) Towards a new definition of multilingual lexicography in the era of internet. In M. J. Domínguez Vázquez, M. Mirazo Balsa, & C. Riveiro Valcárcel (Eds.), *Studies on multilingual lexicography* (pp. 9–28). de Gruyter.
- Haensch, G. (1991) Die mehrsprachigen Wörterbücher und ihre Probleme. In F. J. Hausmann, O. Reichmann, H. E. Wiegand, & L. Zgusta (Eds.), *Wörterbücher: ein internationales Handbuch zur Lexikographie*, 3 (pp. 2909–2937). de Gruyter.
- Hartmann, R. R. K., & James, Gregory (1998) *Dictionary of lexicography*. Routledge.
- Zgusta, L. (1971) *Manual of lexicography*. Academia.

Zita Hollós

Károli Gáspár University of the Reformed Church in Hungary, Budapest

hollos.zita@kre.hu

Busy as a bee!

From a corpus-guided collocations dictionary to a learner's dictionary portal for German

Keywords: collocations, valency, syntagmatic dictionaries, corpus-guided lexicography, learner's dictionary portal, dynamic user interface

The presentation gives a short overview of mono- and bilingual, corpus-guided syntagmatic printed and/or online dictionaries with German and/or Hungarian and at the same time a brief outlook for possible improvements towards a dynamic dictionary portal for language learning purposes with focus on the lexicon.

The first part describes the lexicographic practice of corpus-based valency and collocations dictionaries developed for the two different types of syntagmatic relations: on the grammatical level and on the lexical level (Benson, 1985). We will focus on corpus-based monolingual dictionaries with German, which were published in print and later digitalized.

The second part presents two printed dictionaries which tried to combine both syntagmatic levels, either with a focus on valency or on collocations. The corpus-driven “dictionary” *Hungarian Verb Phrase Constructions* (Sass et al., 2010) is an innovative monolingual product of machine language processing, while the bilingual *KOLLEX – Deutsch-ungarisches KOLlokationsLEXikon* (Hollós, 2023) is some kind of classical lexicographical work compiled in a database with primary and secondary corpus-based lexical sources. The collocations were extracted from the *Leipziger Korpora* (Leipzig Corpora Collection), for the valency realizations the corpus-based *Valenzwörterbuch deutscher Verben* (VALBU) was used.

The last part of my lecture presents the prototype of a learner's dictionary portal *E-KOLLEX DAF* (Hollós, 2016). It illustrates solutions for the integration of corpora, corpus statistic tools and new online lexical resources in the online version of a particular part of the corpus- and data-based bilingual syntagmatic learner's dictionary *KOLLEX* by means of concrete examples. The lecture provides not only insights into the web design and functionalities of this dictionary portal for German as a foreign language, but shows also new developments of the last two years, such as dynamic features of the web design or the so called “Kultmodul”, a new data type for culture specific information, where two possible methods – first a classical then an AI-assisted – are presented to demonstrate everyday challenges of practical lexicographical work in the 21st century.

References:

Benson, M. (1985) Lexical Combinability. *Papers in Linguistics*, 18, 3-15.

Hollós, Z. (2016-) *E-KOLLEX DAF. Elektronisches KOLlokationsLEXikon Deutsch als Fremdsprache. Deutsches Lernerwörterbuchportal für die Sprachproduktion.*
<http://kollex.hu/szotar/>

Hollós, Z. (2023) *KOLLEX: Deutsch-ungarisches KOLlokationsLEXikon. Korpusbasiertes Wörterbuch der Kollokationen Deutsch als Fremdsprache. SZÓKAPTÁR: Német–magyar SZÓkapcsolatTÁR. Korpusalapú kollokációs tanulószótár.* IDS-Verlag.
DOI: <https://doi.org/10.14618/tj9d-7r88>

- Leibniz-Institut für Deutsche Sprache. *Wörterbuch zur Verbvalenz. Grammatisches Informationssystem grammis.* <https://grammis.ids-mannheim.de/verbvalenz>
DOI: <https://doi.org/10.14618/evalbu>
- Sass, B., & Váradi, T., & Pajzs, J., & Kiss, M. (2010) *Magyar igei szerkezetek. A leggyakoribb vonzatok és szókapcsolatok szótára.* [Hungarian Verb Phrase Constructions. Dictionary of the most frequent valencies and word combinations]. Tinta Könyvkiadó.
- Schumacher, H., & Kubczak, J., & Schmidt, R., & de Ruiter, V. (2004) *VALBU - Valenzwörterbuch deutscher Verben.* Gunter Narr Verlag. (Studien zur Deutschen Sprache; 31.)

The closer you look, the more interesting it gets: A Construction Grammar-based analysis of the Slovak Comparative Correlative

Keywords: Comparative Correlative, Slovak, corpus study, Construction Grammar, meso-constructions

The Slovak comparative correlative (CC) is a bi-clausal construction (1):

- (1) [Čím viac čítam,]_{C1} [tým viac rozumiem.]_{C2}
čím more read:I TÝM more understand:I

Each subclause C1 and C2 is introduced by a lexically/phonologically fixed clause-initial element (*čím_{C1}/tým_{C2}*), followed by an obligatory comparative element slot (filled with *viac* in example (1)), followed by a clause slot that does not have to be realized (cf. *Čím ďalej tým lepšie* ‘The further, the better’). Following these observations, Horsch (2021, p. 198) has suggested the following template based on Culicover and Jackendoff’s template for the English CC (1999, p. 567):

- (2) [čím [...]_{comparative element 1 (clause1)}]_{C1} [tým [...]_{comparative element 2 (clause2)}]_{C2}

Regarding its semantics, the CC encodes two meanings: Asymmetric cause-effect ($C1 \rightarrow C2$; ‘reading more results in understanding more’) and symmetric parallel-change-over-time (‘as you read more during a time period x, so you will understand more during a time period y’) (Horsch, 2021, p. 216). With its distinct form (i.e., fixed elements paired with obligatory/optional ‘slots’) and meaning, the CC lends itself well to a Construction Grammar (CxG)-based investigation. CxG assumes that beyond individual words, partly schematic patterns such as the CC can also function as linguistic signs, i.e., conventionalized form-meaning pairings, which are known as constructions (Hoffmann, 2013, p. 310).

In this study, I present a 3,002-token corpus study based on data from the Slovak Web 2011 corpus that explores various formal aspects of the Slovak CC construction, with the goal of contributing to our understanding of this construction. *Inter alia*, I investigate cross-clausal associations that suggest syntactic interdependence between C1 and C2 and therefore, the existence of so-called meso-constructions. Meso-constructions are “semi-productive, partly substantive, partly schematic intermediate meso-constructions” (Hoffmann et al., 2019, 26). I also look at the C2C1 order (e.g. [*filozofi boli tým lepší*],_{C2} better [*čím boli starší*]_{C1} ‘The philosophers were better, the older they were’ (from Horsch, 2021, p. 200)) and variation in slot order, e.g. with the comparative element ‘embedded’ in the clause (e.g. [*čím som sa viac usiloval*]_{C1} [...], ‘the harder I tried...’ (from Horsch, 2021, p. 199)). My results show that while such structures are possible, there is a strong preference for ‘iconic’ alternatives (i.e., C1C2 and with the comparative elements immediately following the clause-initial elements) due to cognitive preferences, specifically, the “Principle of Iconicity” (Bybee, 2012, p. 529), which

posits that the meaning carried by constructions is mirrored in their form.

References

- Bybee, J. L. (2012) Domain-general processes as the basis for grammar. In M. Tallerman & K. R. Gibson (Eds.), *The Oxford Handbook of Language Evolution* (pp. 528–536). Oxford University Press.
- Culicover, P. W., & Jackendoff, R. (1999) The View from the Periphery: The English Comparative Correlative. *Linguistic Inquiry*, 30(4), 543–571.
- Hoffmann, Th. (2013) Abstract Phrasal and Clausal Constructions. In T. Hoffmann & G. Trousdale (Eds.), *The Oxford Handbook of Construction Grammar* (pp. 307–328). Oxford University Press.
- Hoffmann, T., Horsch, J., & Brunner, T. (2019) The More Data, The Better: A Usage-based Account of the English Comparative Correlative Construction. *Cognitive Linguistics*, 30 (1) 1–36.
- Horsch, J. (2021) Slovak Comparative Correlatives: A Usage-based Construction Grammar Account. *Constructions and Frames*, 13 (2) 193–229.

Lana Hudeček, Milica Mihaljević
Institute for the Croatian Language, Zagreb, Croatia
lhudecek@ihjj.hr, mmihalj@ihjj.hr

The Croatian Web Dictionary – Mrežnik – from corpora to dictionary

Keywords: Croatian Web Dictionary, Croatian lexicography, born-digital dictionaries, corpus, corpus based dictionary

The Croatian Web Dictionary – Mrežnik is a growing, born-digital, corpus-based, free, monolingual, easily searchable, hypertext, online, descriptive and normative dictionary of the standard Croatian language (Hudeček, Mihaljević & Jozić, 2024). It is available online in the demo version from A to K (<https://rjecnik.hr/mreznik/>). It consists of three modules – for adult native speakers, for schoolchildren and for non-native speakers of Croatian. Each module is based on a different corpus (or corpora).

The module for schoolchildren is based on the corpus of elementary school textbooks, the module for non-native speakers is based on the corpus of textbooks for Croatian language learners and the module for native speakers on these two online corpora: the Croatian Web Repository – Riznica (<http://riznica.ihjj.hr/index.hr.html>) and the Croatian Web Corpus – hrWaC (<http://nlp.ffzg.hr/resources/corpora/hrwac/>). Data extraction from the corpora is carried out with the web tool SketchEngine, which enables the display of lexeme context through concordances, WordSketches, and WordSketch differences. Since Croatian still does not have a representative corpus of the standard Croatian language, additional specialized corpora were used, all of which are available via Sketch Engine and were compiled at the Institute for the Croatian Language: Coronavirus Corpus (Ostroški Anić & Štrkalj Despot, 2021), Linguistic Corpus (Mihaljević & Marković, 2023, Chapter 3, p. 110), and Gender Corpus (Mihaljević, 2025, Chapter 2, p. 11). Two other new general corpora were also consulted: MaCoCu (Bañón, 2023) and CLASSLA (no-Sketch Engine). The corpora were used for: 1. compiling the headword list, 2. extracting examples and collocations, 3. searching for new meanings not yet recorded in the published Croatian dictionaries, 4. distinguishing meanings and contexts of use.

The presentation will focus on the use of these seven different corpora for different headwords in the *Mrežnik* module for adult native speakers of Croatian and illustrate new meanings, contexts, collocations, and usage notes extracted from the corpora.

References:

- Mihaljević, J. (2025) Rodni korpus i programska rješenja. In: Mihaljević, M. & Hudeček, L. (Eds.) *Muško i žensko u hrvatskome jeziku*. (pp. 11–30). Institut za hrvatski jezik i jezikoslovlje.
- Mihaljević, J. & Marković, M. (2023) Jezikoslovni korpus. In: Mihaljević, M., Hudeček, L., Jozić, Ž. (Eds.) *Hrvatsko jezikoslovno nazivlje*. (pp. 110–123). Institut za hrvatski jezik i jezikoslovlje.
- Hudeček, L., Mihaljević, M. & Jozić, Ž. (Eds.). (2024) *Anatomija rječnika. Hrvatski mrežni rječnik – Mrežnik*. Institut za hrvatski jezik. Zagreb.
- Bañón, M.; et al. (2023) *Croatian web corpus MaCoCu-hr 2.0*, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042. <http://hdl.handle.net/11356/1806>

Štrkalj Despot, K. & Ostroški Anić, A. (2021) A War on War Metaphor: Metaphorical Framings in Croatian Discourse on Covid-19. *Rasprave Instituta za hrvatski jezik i jezikoslovlje*, 47 (1) 173–208. <https://doi.org/10.31724/rihjj.47.1.6>

Kathryn M. Hudson
Eckerd College on Florida's Gulf Coast, Florida, USA
hudsonkm@eckerd.edu

Re-Rooting the Self: Historical Lexicography and Identity in Contemporary Mesoamerica

This paper uses biolexicography and critical lexicographic perspectives to explore how historical lexicographic materials contribute to constructions of indigenous identity in contemporary Central America by facilitating (re)engagement with the natural world in ways that (i) offer a sense of legitimacy through connections with a recognized and documented history and (ii) facilitate claims of cultural and linguistic competence. The original cultural source and precise referent of recorded lexemes are less significant than their occurrence in a historical lexicographic record – an inclusion that symbolizes historical validity and indigeneity for many modern populations. These materials thus function as sources of cultural, historical, and linguistic information – most commonly in relation to faunal, floral, and topographic referents – that can be used to develop and/or assert indigenous identities. Engagements with these documents and their lexicographic contents are constructed and conceptualized as proof of identity and as a tangible connection to a past that has been largely erased. These functions are particularly significant in Central America, where the effects of colonial and postcolonial histories on indigenous identities have been especially dramatic. Critically considering the role of historical lexicographic data in mitigating these effects illustrates the centrality of ecological connections in the (re)establishment and (re)assertion of identity. It also provides insights into the emergence of two distinct but closely related categories of identity – a generalized pan-indigenous identity and more specific local or regional identities connected to historically documented groups – and their rooting in the natural world.

Katarzyna Kryńska
Institute of Polish Language Polish Academy of Sciences, Warsaw
katarzyna.krynska@ijppan.pl

Implementing Greek into OCR Model of Multilingual Old Dictionary

Keywords: Lexicography, Greek, OCR, Training Models

The aim of the presentation is to discuss the process of implementing Greek letters during the training of an OCR model marked for text recognition in the multilingual Polish-Latin-Greek dictionary by Grzegorz Knapiusz (1643).

The digital edition of Knapiusz's *Thesaurus polono-latino-graecus* is one of the stages of the project aimed at creating the Database of Historical Polish Lexicons and implemented at the Institute of Polish Language, Polish Academy of Sciences. The project also provides for the digital edition of two other dictionaries.

Particular problems arose at the stage of implementing Greek when training the model for the Knapiusz's dictionary. The Greek text of the dictionary required introducing into the model another alphabet with all its special characters. The text contains many ligatures and abbreviations that the model had to learn. We are able to develop them thanks to William Wallace's "Index of Greek Ligatures and Contractions" (1923). Moreover, the dictionary often contains variant spellings of letters, ligatures or abbreviations, which required the model to remember that a given word or letter can be written in different ways. There are probably more ligatures and abbreviations in the entire dictionary than in the sample of material used to train the model, so the proofreading team will play a key role once the OCR work is completed. Another challenge was the diacritics of the Greek language: accent marks, breathing marks, apostrophes and iota subscriptum due to the variety of characters within the same letter and the multiplicity of characters within one letter or in close proximity. These marks were not recognized correctly for a long time during the model training process. The process of learning the OCR model to recognize a Greek text, with its problems and solutions, will be presented on the poster.

Fig. 1. Greek ligatures and abbreviations in the dictionary by Knapiusz (1643), examples

❖ Greek ligatures and abbreviations (examples):

Ϡ	ος	βίϠ	ἐρίϠϠ	ἰω	ην	γλω̣ν	ἀναίδῳ	ἰω:		
Ϛ	στ	ὄστρακα	ἰστάμενος	κίση	χ ^ο	σχο	χολάζω	χολή		
Ϙ	ου	μϘ	τϘ	χρυσὸν	κρυψίνης	ω̣	σθαι	κεῖω̣	κεῖω̣	γένεω̣
πω	ππ	ἰπω	ικὸς							

Prepositions

ὑπο	ὑπο	ὑποδίξω	ὑποπνέω	ὑποτείνω	περί	περί	περίπτερϠ	περίσυλον	πεικοσμέω
παρά	παρά	παράνοια	παράφρασις	παράγω	διὰ	διὰ	διὰτέλλω	διὰτεταγμένοι	

Whole words

τὸν	τῆς	ταῖς
τούς	καὶ	

Fig. 2. Variant notation of Greek letters, ligatures and abbreviations

❖ Variant notation of Greek letters, ligatures and abbreviations

β ἐκβιβαστῆς τὸ μολίβδωμα, ἡ μολίβδαινα

π πηλακῶν πεπληρομένῳ πάντας

τ ψιλότης διαφράγματι περὶ τῷ

κατὰ κτ̃ καὶ καὶ σθαι σθ̃ ος ϑ̃

Fig. 3. Variety and multiplicity of diacritical marks

❖ Variety and multiplicity of diacritical marks

ὦ ὦμιωρ ἔω ἦ ἦ εἴτε εἶναι ἐρρῶμιωρῶρ ἀπ' ἀνθρῶρ

Anna Lehocki-Samardžić¹, Szilvia Szoták²

¹Josip Strossmayer University, Osijek, Croatia

²Budapest Metropolitan University, Budapest, Hungary

analehocki@gmail.com, sszotak@metropolitan.hu

Examining the educational terms of the Termini Hungarian-Hungarian dictionary

Keywords: Termini Research Network of the Hungarian Language, Termini Hungarian-Hungarian e-dictionary, contact elements in Hungarian language, 'de-convergence', educational terms, country-specific term, database, lexicography

The main research project of the Termini Hungarian Language Research Network is the construction and expansion of the Termini Hungarian-Hungarian e-dictionary. The dictionary contains state language contact elements and non-standard Hungarian words used by the Hungarian minority living in neighbouring countries. The aim of the research project is to familiarise Hungarians in Hungary with the contact elements of the official language used by Hungarians living abroad in their everyday communication. As Hungarian is a pluricentric language, the significance of the dictionary lies in the fact that it will help to better understand the linguistic varieties of the different language regions.

The emergence of terms that differ from those used in Hungary began with the border changes after 1920, which resulted in a process of linguistic fragmentation. This process of linguistic divergence was replaced by convergence after 1989, thanks to political, economic, cultural and scientific changes. The members of the Termini Research Network *refer to* this process *as 'de- convergence'*.

In our presentation, we will examine the educational terms in the Termini Hungarian-Hungarian e-dictionary used by Hungarians in Croatia and Austria in relation to the educational systems of the two countries. When expanding the dictionary, we follow the uniform lexicographical methodology manual developed by the Termini Research Network (Lanstyák, Benő, Juhász 2010). We argue that while the process of divergence in the language use of Hungarian ethnic groups in the neighbouring countries has been replaced by convergence, a number of diverging factors can be observed in the changes in the educational policies and the development of the educational systems of the neighbouring countries over the last 100 years (Lehocki-Szoták 2022). In order to examine divergences, it is necessary to study and compare the educational systems of the countries mentioned, with that of Hungary. The interpretation of the term is aided by the methodology, i.e. the definition, the number of meanings, the conceptual classification (currently 46), the name of the region/geographical language variant, and the determination of style classifications (which may refer to dialect, register, style variation, temporality, emotional attitude, and relative frequency).

In this presentation we will illustrate these differences through the terms of the dictionary. We will explore the factors that help us to understand how divergent educational policy processes on different sides of the border have reinforced divergence.

References:

- Lanstyák I., Benő A., és Juhász T. (2010) A Termini magyar–magyar szótár és adatbázis. *Regio* 21 (3) 37–59.
- Lehocki-Samardžić A., Szoták Sz. (2022) Oktatási terminusok és oktatási rendszerek: avagy mi konvergál, és mi divergál? In: Karmacsi Z., Márku A., & Máté R. szerk. *A határ mint konvergáló és divergáló tényező a nyelvben. Tanulmányok a 21. Élőnyelvi Konferenciáról*. Törökbálint: II. Rákóczi Ferenc Kárpátaljai Magyar Főiskola Hodinka Antal Nyelvészeti Kutatóközpont–Termini Egyesület 133-143.

Miguel Angel Lopez, Elena Andreeva, Larissa Ferreira

Universität Hildesheim, Germany

aurelia0770@gmail.com, ferreiral@uni-hildesheim.de, lopezm@uni-hildesheim.de

Ausgewanderte Wörter: An online Dictionary of German Emigrant Words

Keywords: Emigrant words, Loanwords, Contrastive lexicography, Multilingual dictionary, Language contact, Cultural transfer, Digital lexicography

Emigrant words (*ausgewanderte Wörter*) represent a significant linguistic phenomenon in contrastive lexicography, reflecting processes of language contact and cultural transfer. In this study, we present a multilingual online dictionary that documents and analyzes words of German origin that have been incorporated into Spanish, Portuguese, Russian and English. The dictionary is developed on a digital platform using a MySQL database as well as PHP and Javascript, allowing for dynamic navigation between terms and their equivalents in different languages. Lexicographic relationships between words have been established, considering aspects such as phonetic and morphological adaptation, usage domains, and frequency in each target language.

From a methodological perspective, we combine corpus analysis, previous lexicographic studies, and a semantic classification of the terms. Additionally, we explore the linguistic migration paths of these words, identifying historical and cultural patterns that have facilitated their transfer.

This dictionary not only contributes to the study of loanwords from a lexicographic perspective but also provides an accessible tool for researchers and speakers interested in the influence of German on other languages. In addition, we will discuss the technical and conceptual challenges of the project, as well as its potential future developments.

Daniela Majchráková

Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences, Bratislava, Slovakia

majchrakovadaniela@gmail.com

Compiling the Slovak Dictionary of Adverbial Collocations

Keywords: dictionary of collocations, Slovak adverbs, collocability, collocational structure, balanced corpus

The Slovak dictionary of adverbial collocations that is currently being created is a continuation of the previous two dictionaries capturing the collocability of Slovak nouns and adjectives. In its final version, the Slovak dictionary of adverbial collocations will contain collocational profiles of more than 700 most frequent adverbs of the contemporary Slovak language. The dictionary will also provide statistical information on adverbs and their collocates, and for this purpose a special balanced corpus based on the data from Slovak National Corpus has been created.

The collocational profiles consist of collocates, words that occur most frequently in Slovak texts with adverbs and create typical or fixed word combinations with them, e.g. *tuho premýšľať/think hard*, *kolmo nadol/straight down*, *štedro odmenený/generously rewarded*. Primarily, it is a registration of statistically significant collocates, verbal, adjectival, adverbial, or particle collocates, which create mainly typical word combinations.

The analysis of the collocational potential of adverbs shows that in addition to the basic collocational structures mentioned above, there are also various structural types, many of which are typical of semantically bounded/limited groups of adverbs. For example collocations with adverbs of time also form combinations with synsemantic words such as prepositions or conjunctions, e.g. *dnes a zajtra/today and tomorrow*, *rovnako ako minule/just as before*, *v noci na včera/on the night before yesterday*, *vlani v lete/last summer*. The analysis also shows that the adverbial collocational profile is not only a special set of lexical units, but that its uniqueness lies in the way the adverb is associated with these units.

The dictionary is planned to provide insight into the semantic and morphosyntactic relationships of adverbs with other word types. The publication will be of benefit both for the common user and the didactics of Slovak language teaching and will provide an important material base for possible further in-depth linguistic research on adverbial collocability.

Magdalena Majdak

Institute of Polish Language, Polish Academy of Sciences, Warsaw, Poland

[magdalena.majdak@ijppan.pl](mailto:magdalenamajdak@ijppan.pl)

The Database of Historical Polish Lexicons: Concept and Lexicographic Challenges¹

Keywords: Database of Historical Polish Lexicons, historical dictionaries, metalexicography, headword, definition

The poster presents the Database of Historical Polish Lexicons, a project (2024–2029) funded by the National Program for the Development of the Humanities and carried out at the Institute of the Polish Language, PAS.

The project involves the digital edition of three key 17th- and 18th-century Polish dictionaries: *Thesaurus polono-latino-graecus* (1643) by Knapiesz, *Nowy dykcjonarz* (1764) by Troc, and *Forytarz* (1674) by Ernesti, which is organized in a tergo order. These dictionaries represent an early stage of lexicographic development, each characterized by a unique arrangement of material, a distinct structure of dictionary entries, and various unresolved editorial issues.

Creating the database requires not only preparing a digital edition of the dictionaries but also analyzing their microstructure and macrostructure, and above all, developing a methodology for comparing dictionary entries and their content across dictionaries (both within this project and with future additions). An important benefit will be the first systematic access to Polish lexical material, enabling the extraction of words that appear in different parts of dictionary entries.

The poster will present the following technical and lexicographic topics:

- Presentation of the database schema on the TEI Publisher platform,
- Search capabilities within and across dictionaries,
- Displaying dictionary entries with cross-references to other dictionaries and ideas for their alignment,
- The challenge of identifying headwords in historical dictionaries, for example:
 - a) Up to the beginning of the foreign-language text – e.g., *Szyszka miękka podługowata iako ogonek na drzewach niektórych / iako leszczynie*.
 - b) Up to the first slash or period – e.g., *Dzierżawa/ maiętność którą kto dzierży* but also *Skrzeczq/ Rzegocq żaby* or *Dzieci ptasze/ rybie/ gadziny/ y innych zwierząt*.
 - c) The first word or a prepositional phrase – but also *Dzieci opatrujący*.

Discussing the project internationally will provide valuable insights for developing historical multilingual dictionary databases.

¹ The research was carried out as part of the project ‘Third-level digitization of large 17th and 18th century dictionaries: Creation of a database of historical Polish lexicons’ (no. NPRH/DN/SP/0003/2023/12) financed by the Ministry of Science and Higher Education.

Susana Duarte Martins
NOVA CLUNL - Linguistics Research Centre of NOVA University of Lisbon
NOVA FCSH - Faculty of Social and Human Sciences of NOVA University of Lisbon
susanaduartemartins@fcsb.unl.pt

Educational applications of general language dictionaries in advanced language proficiency

Keywords: digital dictionaries; educational resources; language teaching; lexiculture; Portuguese as a foreign language

Given their pedagogical merit, dictionaries are recommended instructional materials for language teaching and are recognized by many methodological approaches. However, these lexicographic resources are often overlooked in advanced language learning, as evidenced by international language standards such as the *Common European Framework of Reference for Languages* (CEFR) and its *Companion Volume* (Council of Europe 2001, 2020). Consequently, the full potential of dictionaries as tools for advanced language proficiency remains untapped, with their pedagogical use often relegated to more elementary levels of language teaching (Author 2023).

This talk addresses the contributions of general language dictionaries for proficient users and highlights the benefits of their use in language teaching. Particular focus is given to digital dictionaries as educational tools that support the development of spoken and written comprehension and production activities in the foreign language classroom. Case discussions demonstrate the crucial role of dictionaries in fostering linguistic knowledge among international students enrolled in advanced -level courses of European Portuguese as a foreign language.

The examination of the macrostructure and microstructure of selected lexical entries from the three leading digital dictionaries of European Portuguese (*Dicionário de Língua Portuguesa* da Academia das Ciências de Lisboa, *Infopédia*, and *Priberam*) illustrates how these resources contribute to the acquisition of new knowledge. This includes the incorporation of language usage marks, idiomatic expressions, slang, and other types of information that provide access to cultural elements shared by native speakers (Galisson 1987), which are especially significant in pluricentric languages such as Portuguese.

Lastly, the educational applications of digital dictionaries stimulate learner -centered approaches by enhancing students' active participation in their learning process. Simultaneously, these tasks support the development of lexicographic and digital literacy skills within multilingual and multicultural contexts.

References:

- Academia das Ciências de Lisboa (2023) *Dicionário da Língua Portuguesa* (A. Salgado, Coord.). Academia das Ciências de Lisboa /Institute of Lexicology and Lexicography of the Portuguese Language. <https://dicionario.acad-ciencias.pt/>
- Council of Europe (2001) *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge University Press.
<https://rm.coe.int/16802fc1bf>
- Council of Europe (2020) *Common European Framework of Reference for Languages: Learning, teaching, assessment – Companion volume*. Council of Europe Publishing.

www.coe.int/lang-cefr

Duarte Martins, S. (2023) Digital dictionaries as a pedagogical tool in foreign language didactics. *Cahiers de Lexicologie*, 122 (1) 209–238. <https://doi.org/10.48611/isbn.978-2-406-15055-8.p.0209>

Galisson, R. (1987) Accéder à la culture partagée par l'entremise des mots à C.C.P. *Études de Linguistique Appliquée*, 67, 119–140.

Porto Editora (2003–2025) *Dicionário Infopédia da Língua Portuguesa*.
<https://www.infopedia.pt/>

Priberam (2008–2025) *Dicionário Priberam da Língua Portuguesa*.
<https://dicionario.priberam.org/>

Maja Matijević
Institute for the Croatian Language, Croatia
mmatijevic@ihjj.hr

Autohyponymy and Automeronymy in the Croatian Language

Keywords: lexical relations, automeronymy, autohyponymy, vertical polysemy, dictionary definition

Studies on lexical-semantic relations typically contrast the meanings of different lexemes, i.e., formally distinct lexical units. For example, "day" and "night" are antonyms, "animal" and "dog" are a hyperonym and a hyponym, while "body" and "hand" form a holonym-meronym relationship. However, lexical-semantic relations can also exist within a single polysemous lexical unit. In the literature, the most widely studied phenomenon is enantiosemy or autoantonymy, as seen in the case of the Croatian verb "posuditi", which can mean both 'to borrow' and 'to lend,' thereby forming an antonymous relationship within the same lexeme. This paper examines autohyponymy and automeronymy in the Croatian language and demonstrates how an analysis of a corpus composed of definitions from a monolingual dictionary can lead to the (semi-)automatic identification of autohyponyms and automeronyms. The study presents the methodology for compiling a corpus of dictionary definitions and the targeted search procedure used to analyze the extracted examples. The primary lexical resource used was Croatian Web Dictionary – Mrežnik, the first Croatian web dictionary, and definitions from this source were compared against those found in other monolingual dictionaries of the Croatian language.

The analysis suggests that autohyponymy in Croatian most commonly occurs in lexemes referring to plant and animal species, while automeronymy is primarily found in lexemes denoting plants and their parts. The results highlight the importance of lexicographic sources in lexicological and semantic research.

Nathalie Mederake, Nico Urbach

Marburg University, Research Center Deutscher Sprachatlas, Germany

nathalie.mederake@uni-marburg.de, nico.urbach@uni-marburg.de

Towards a broader view of regional dialects. Cross-linguistic connections in dialect dictionaries

Keywords: lexicography, regional dialects, cross-linguistics, dialect dictionaries

Dialect dictionaries have the task of describing the dialectal vocabulary of an area below the overall national area of a language. But whereas dialect dictionaries mostly concentrate on administrative or territorial divisions dialect areas go beyond the borders of political entities. In fact, one dialect or language variety merges into another and regional variation exists along a dialect continuum rather than as having sharp breaks from one region to the next. The Mecklenburg Dictionary (= MWb) and the Hesse-Nassau Dictionary (= HNWb) are two large-scale dictionaries that document the regional language as unique works. Whereas the MWb covers a large part of the Low German vernacular, the material of the HNWb has shares in the three major linguistic landscapes: Low German, High German and Upper German.

This presentation reports partial results from a larger pilot study that examines the feasibility of establishing links between such dictionaries across language areas and at different levels (e.g., sense and subject-domain). We are in the process of finalizing the datasets, which will serve as the basis for exploring several strategies for linking content; here we focus on a subject-group approach that is gaining traction in current research in dialect lexicography. Using a hierarchical scheme developed to encode subject groups, each word sense is assigned a four-digit code that descends from general to specific domains. Applied consistently to both dictionaries, this labeling enables systematic, onomasiological access to their data and supports cross-dictionary queries that are not constrained by established boundaries. Our preliminary observations suggest that subject-group coding helps reveal how regional semantics is structured and how it differentiates across the wider language area. In the long term, this systematic evaluation provides a foundation for a network of interoperable dialect dictionaries, offering users concept-based navigation and new ways to interact with word information across regions and language varieties.

Peter Meyer
Leibniz Institute for the German Language, Germany
meyer@ids-mannheim.de

Walking the graph: Finding etymological information in a web platform on German loanwords in other languages

Keywords: loanword lexicography, graph database, Linked Data, TEI, Large Language Models

The Lehnwortportal Deutsch (henceforth, LWPD; <https://lehnwortportal.ids-mannheim.de>) is a web platform that provides free access to currently 16 rather heterogeneous resources on German lexical borrowings in other, mostly European, languages. The recently launched new version 2.0 supersedes an older and considerably smaller prototype. Apart from a unified lemma-based access to individual entries in the included dictionaries, the system features a graph-based data architecture representing lexical units (etyma, loanwords, derivatives thereof, ...) and the network of their interrelations across all resources. In order to leverage this data model, the platform offers a visual 'query builder' type of user interface. This enables a simultaneous search for multiple words, each with any number of user-specified properties, standing in specific relationships (e.g. borrowed as, same/different POS/language, etc.) to each other. Throughout LWPD's web application, the network-like structure of the data is emphasized. Naturally, this structure and appropriate advanced search options pose specific challenges to end users. In our presentation, we discuss both the data model and the corresponding user interface from the practical perspective of a non-expert user. We compare the 'labeled property graph' model of the Lehnwortportal to Linked Data approaches (cf. Khan 2018) and TEI-based models (e.g., Bowers et al. 2022) for etymological data; and we contrast the portal's query builder with full-blown visual exploration systems for Linked Data (cf. Jacksi et al. 2018 for a survey). In addition, we discuss lexicological limitations of the graph approach arising, e.g., in the treatment of internationalisms, from the end user's point of view. Finally, we report on experiments to overcome the unavoidable complexity of advanced graph database searches by offering an LLM-driven free-text search option that 'translates' user input into valid query builder configurations which can subsequently be further adapted by the user.

References

- Bowers, J., Herold, A., Romary, L., & Tasovac, T. (2022) TEI Lex-0 Etym – Towards terse recommendations for the encoding of etymological information. *Journal of the Text Encoding Initiative*. <https://doi.org/10.4000/jtei.4300>
- Jaksi, K., Zeebaree, S. R. M., & Dimililer, N. (2018) LOD Explorer: Presenting the web of data. *International Journal of Advanced Computer Science and Applications* 9 (1) 45–51. <https://doi.org/10.14569/IJACSA.2018.090107>
- Khan, A. F. (2018) Towards the representation of etymological data on the semantic web. *Information*, 9 (12) Article 304. <https://doi.org/10.3390/info9120304>

Ana Mihaljević, Josip Mihaljević
Old Church Slavonic Institute, Zagreb, Croatia
amihaljevic@stin.hr, jmihaljevic@stin.hr

The second phase of the development of the online version of the *Dictionary of the Croatian Redaction of Church Slavonic* – New challenges

Keywords: Croatian Church Slavonic, Glagolitic script, e-dictionary, OCR, retro-digitization

The *Dictionary of the Croatian Redaction of Church Slavonic* (RCJHR 2000–) is the first lexicographical description of the Croatian Church Slavonic language. Its compilation is a fundamental and ongoing project of the Old Church Slavonic Institute in Zagreb. The dictionary has been published in printed form from A to I and in recent years intensive research has been done on the development of the online version of this dictionary.

This dictionary has a very complex microstructure with entries in four scripts (Latin, Greek, Glagolitic, Old Cyrillic) and five languages (Croatian Church Slavonic, Croatian, English, Latin, Greek). The first 12 fascicles exist only as scanned PDFs, fascicles 13–17 are available to project members as DOC files with a partially illegible Greek text, while fascicles 18–21 are available as original DOC files.

In the first phase of retro-digitization, a special web portal was set up to display the dictionary and all headwords were entered. To ensure correct display, the system was enhanced with FSGLA and Cyrillic fonts and equipped with a virtual Glagolitic keyboard to facilitate text entry. Each headword was linked to the corresponding PDF page, completing the first phase as described in Mihaljević & Mihaljević (2024a, 2024b).

The second phase focuses on the development of a structured format for entering dictionary entries that facilitates headword search. At the same time, various tools are used to create digital versions of dictionary entries. An important step is the use of Transkribus to develop an OCR model that can recognize the full text of dictionary entries. This model will be trained on the newer fascicles, which are in their original DOC format, and applied to older fascicles.

References:

- Mihaljević, A., & Mihaljević, J. (2024a) Mrežna inačica *Rječnika crkvenoslavenskoga jezika hrvatske redakcije*. *Slovo*, 74, 169–194.
- Mihaljević, A., & Mihaljević, J. (2024b) Digitalizacija i igrifikacija *Rječnika crkvenoslavenskoga jezika hrvatske redakcije*. In S. Marjanović (Ed.), *Zbornik skupa Leksikografski susreti*. Beograd: Filološki fakultet, Univerzitet u Beogradu.
- Staroslavenski institut (2000) *Rječnik crkvenoslavenskoga jezika hrvatske redakcije: A – VRÊD*. Zagreb: Staroslavenski institut.
- Staroslavenski institut (2015) *Rječnik crkvenoslavenskoga jezika hrvatske redakcije: VRÊDbNb1 – ZAPOVÊDNICA*. Zagreb: Staroslavenski institut.
- Staroslavenski institut (2016) *Rječnik crkvenoslavenskoga jezika hrvatske redakcije: ZAPOVÊDb – ZEMLbNb*. Zagreb: Staroslavenski institut.

Sylvain Monnoukoun

Université d'Abomey-Calavi, Département d'Etudes Germaniques, Bénin

sylvmonnouk@gmail.com

Research into Dictionary use: User Behaviors and Preferences in Benin

Keywords: dictionary use, user preferences and expectations, lexicography, digital resources

This study investigates dictionary usage behaviors and preferences among different user groups in Benin, West Africa, focusing on students, teachers, and professionals. The primary objective is to examine how these users interact with dictionaries, the difficulties they encounter, and their preferred features and formats. The research employs an online survey distributed via Google Forms to gather comprehensive data from 300 participants across educational and professional settings. The survey includes questions about the frequency of dictionary use, preferred formats (print, digital, mobile), and the impact of Benin's linguistic diversity on accessibility. The survey examines dictionary users' behaviors, needs, and expectations, while interviews explore their usage skills. It examines dictionary users' behaviors, needs, and expectations, while semi-structured interviews with 30 selected respondents provide deeper insights into their experiences and challenges with dictionary usage. Additionally, semi-structured interviews are conducted with a select group of respondents to gain deeper insight into their experiences and challenges with dictionary usage. Key obstacles such as limited digital access, lack of training, and user expectations are highlighted. Expected outcomes include a comprehensive profile of dictionary usage patterns, insights into the most valued features—such as definitions, examples of usage, and audio pronunciations—and the identification of barriers to effective use. Ultimately, the findings inform the design of lexicographic resources that are more accessible and relevant to the needs of users in Benin, enhancing their language learning and communication experiences. In summary, this study provides valuable insights into dictionary usage in Benin, paving the way for the development of tailored resources that meet the diverse needs of its users. By employing a methodology that combines online surveys and interviews, the research ensures a comprehensive understanding of user behaviors and preferences, facilitating the creation of more effective and user-friendly dictionary resources.

References:

- Atkins, B.T. Sue and Varantola, Krista. (1998) Monitoring Dictionary Use. In Atkins B.T. Sue (ed.) 1998: 21–82.
- Bergenholtz, H., & Tarp, S. (2003) *User-oriented lexicography: Theoretical and practical perspectives*. In *Lexicography: Principles and practice* (pp. 1–20). London: Routledge.
- Deschryver, G. (2009) "Access Structures in Electronic Dictionaries: A User- Centered Approach." In *Electronic Lexicography in the 21st Century: New Challenges, New*

- Applications*, edited by A. P. Cowie, 123–134. Oxford: Oxford University Press.
- Hartmann, Reinhard R.K. (2013) Aids in Metalexicographic Research. Gouws, Rufus H., Ulrich Heid, Wolfgang Schweickard and Herbert Ernst Wiegand (Eds.). 2013: 596–611.
- Wiegand, H. E. (2003) Wörterbuch zur Lexikographie und Wörterbuchforschung (WLWF)/ Dictionary of Lexicography and Dictionary Research. Eine kurze Projektbeschreibung. In: Botha, W.F. (ed.) 2003. n Man wat beur. Buro van die WAT: Stellenbosch: 368-384.

Perina Vukša Nahod, Bruno Nahod
Institute for the Croatian Language, Zagreb, Croatia
pvuksa@ihjj.hr, bnahod@ihjj.hr

From apron to fork – a new approach to Croatian dialectological dictionary

Keywords: dialectological dictionary, Čakavian, Kajkavian, Štokavian, lexicography

Croatian dialect lexicography is enriched each year with numerous amateur and scholarly publications, and it is extremely important for gaining insight into the material from individual local dialects or multiple speech varieties. The creation of dialect dictionaries is challenging because it requires devising a precise and comprehensive structure for each dictionary entry. There are two fundamental problems in Croatian dialect lexicography: the lack of an established computer program for dictionary compilation and the online inaccessibility of the published material.

Until now, there has been no thematic dialectological dictionary that covers all Croatian dialects. Therefore, our goal was to process dialect lexemes from the semantic field of kitchen (which encompasses areas such as fruit, vegetables, utensils, furniture, etc.). Although many dialect dictionaries exist in various formats, we did not have a model for creating the envisioned dictionary that would focus on just one thematic field: kitchen, and display material from all Croatian dialects: Čakavian, Kajkavian, and Štokavian.

Given our objective to ensure that all data is both searchable and accessible, we have elected to compile the dictionary using the lexicographic software TLex. Employing this professional tool guarantees a systematic and consistent editing of the data, thereby streamlining its preparation for online publication.

This paper will present the method for entries' main structure, which includes grammatical forms, phraseological, etymological, and oral literature processing, geolocation, phonetic transcription, speech sample, photograph, map, and literature. It will also showcase possible solutions for linking dialect material with standard language and terminological dictionaries. The corpus consists of material available from published and manuscript dialect sources, as well as data obtained from independent field research.

With this approach, we aim to gather more linguistic data in one place, modernize the input and presentation of Croatian dialect material, and ultimately increase its visibility.

Barbara Patella
Università degli Studi di Firenze, Italy
barbara.patella@unifi.it

**Digitize dictionaries using XML-TEI: a vademecum of methodologies and applications
(based on non-alphabetic resources)**

Keywords: electronic dictionaries, non-alphabetical dictionaries, XML-TEI, lexicography, history of the Italian language

This paper aims to provide, in the field of digital lexicography, a vademecum for the markup of retro- digitized dictionaries, a sort of theoretical-practical handbook containing methodologies and XML-TEI tagging solutions functional to their transposition (and enhancement for linguistic studies) from paper to electronic format. The topic will particularly focus on non-alphabetic dictionaries, therefore on onomasiological resources, which represent a real test-bed for the mark-up activity. The operations foreseen and foreseeable for dictionaries digitization will be reviewed: from the study of the lexicographic structure to the definition of the tagset (based on XML-TEI *Guidelines*), from the application phases to the design of the search modes (FIGG. 1-2), culminating in the setting up of queryable platforms that will host the resources in electronic version (FIG. 3). In short, we will show the entire workflow that a linguist (with some computer skills) must manage when working on electronic lexicography projects.

The essay originates from two case studies related to the Italian language: the creation of an electronic version of nine 19th century non-alphabetic dictionaries and one of the *Vocabolario toscano dell'arte del disegno* by Filippo Baldinucci, a 17th century specialised dictionary on the language of art – in the matter in question, retro- digitized repertoires marked up using *Oxygen XML Editor*.

We therefore begin with the digitization of these resources with the aim of proposing general solutions, that can be applied to similar lexical or lexicographic works (such as glossaries, encyclopedias, collections of proverbs, etc.). Additionally, we propose to reflect on methodological challenges in the panorama of electronic lexicography and present concrete and usable results in the field of digital humanities, highlighting the importance of analysis and tagging levels, which can be modulated and customized depending on the specific research purposes, timing (thus the sustainability and feasibility of the project) and end-use of the digital product.

STRUMENTI A CORDE, denominazione generica di quelli strumenti, nei quali il suono è prodotto dalle vibrazioni di Corde tese. Codede Corde sono o di minugia, cioè di budello, o d'un filo d'ottone, oppure anche **FASCIATE**: queste ultime si formano coll'avvolgere spiralmente un sottil filo metallico in giri stretti e contigui intorno a un'ANIMA, la quale ora è essa stessa una Corda metallica: come nelle Corde basse del Cembalo, ora è una Corda di legno, come nella Quarta del Violino, ora è di più fili, come nelle tre più gravi Corde della Chitarra, e in parecchie dell'Arpa.

Nota 16. Negli strumenti a corde l'acutezza del suono è in ragione diretta della tensione e della corda, e inversa del diametro, e della lunghezza della medesima. Vale a dire che l'acutezza del suono cresce col crescere la tensione della corda, e cresce col diminuire la grossezza e la lunghezza della medesima: in altri termini, date due corde ugualmente grosse, e lunghe, ma inegualmente tese, la più tesa darà suono più acuto: date due corde ugualmente tese, e ugualmente grosse, ma inegualmente lunghe, la corda meno lunga darà un suono più acuto: finalmente a parità di tensione e di lunghezza, il suono più acuto sarà dato dalla corda meno or

Questo teorema non è se non un semplice corollario del generale teorema esposto nella Nota 9, il quale insegna che in qualsiasi strumento il suono più acuto è prodotto da un maggior numero di vibrazioni fatte in un dato tempo. E questo maggior numero di vibrazioni le fa appunto la corda più tesa, la corda meno grossa, e la corda meno lunga.

La tensione delle corde, che negli strumenti a corda si misura in iperonimo, si riferisce al grado di tensione delle corde, che influisce sulla loro vibrazione e sulla produzione del suono. La tensione delle corde è regolata dai meccanismi di sintonizzazione dello strumento, che permettono di adattare la tensione delle corde alle diverse note della scala musicale.

Gli strumenti a corde, secondo i varj modi di eccitarne le vibrazioni sonore, possono dividersi nelle seguenti specie:

STRUMENTI A CORDE PICCHIATE cioè percossi da un corpo sodo. Tali sono, per es., il Cembalo e il Salterio.

CÈMBALO, strumento a corde metalliche, per lo più orizzontali, ta-

lora verticali, tese mediante i **PIRONI**, cioè cilindretti di ferro a testa compressa, volgendo i quali con una chiave si tendono le corde, fatte suonare da altrettanti pezzi di grosso filo metallico, ripiegati in forma di gancio, a guisa di martellini cadenti, mossi col toccare i varj **TASTI** della **TASTIERA**, o anche col premere co' piedi alcuni **PEDALI**.

Nota 17. Questo strumento fu già chiamato GRANTICEMBALLO, CLAVICEMBALLO, BUONACCORDO, ARPICORDO, e dai moderni, nello scorcio del passato secolo, cominciò a chiamarsi PIANOFORTE (V. Dizionario o Vocabolario universale di Napoli, ad vocem), le quali variate denominazioni forse accennarono già ad alcuni successivi miglioramenti, che vi si andarono facendo.

Ma qui non è da tacersi, che gli antichi Classici Italiani danno unicamente il nome di *Cembalo* a un rozzo strumento, che più avanti sarà da me registrato. (V. TAMBURELLO, e Nota 23).

SPINETTA sorta di piccol Cembalo, che ha un minor numero di corde, e queste sono scosse da altrettanti pezzi di cannello di penna di corvo, tagliata in punta, e infissa nei **SALTARELLI**, mossi dai corrispondenti tasti che il sonatore tocca colle dita.

SALTERO, cassa di legno sottile, in forma di trapezio, alta circa un sommosso, lunga e larga circa tre palmi, sul Coperchio della quale sono parecchie corde metalliche, sostenute in piano da PONTICELLI laterali, e tese da altrettanti PINONI di ferro, piantati lungo i due lati non paralleli.

Ogni suono è renduto da tre, anche da quattro corde unisone.

Suonasi picchiando con due piccole BACCHETTE di legno, una per ciascuna mano: ovvero coi DITALI, che sono come altrettanti anelli da cucire, incastrati in cima delle dita delle mani, in ciascuno de' quali è fermato un pezzo appuntato di canello di penna: anche suonasi a mano, cioè col nudo polpastrello delle dita.

Nota 48. Pare che non metta conto di separare dai predetti strumenti a corde picchiate la MANDOLÀ e 'l MANDOLINO, nei quali le corde, anzi che picchiate propriamente, sono più tosto scosse con un pezzetto di cosa sottile, cedevole, elastica, che chiamano la PENNA, sia essa veramente una penna, oppure una laminetta assottigliata di tartaruga, ovvero di scorza elastica di ciliegio.

FIG. 1: An example of the structure of Carena's *Vocabolario d'arti e mestieri* (1853, pp. 40–41): the relation between hypernyms and hyponyms (starting from the lemma *strumenti a corde*).

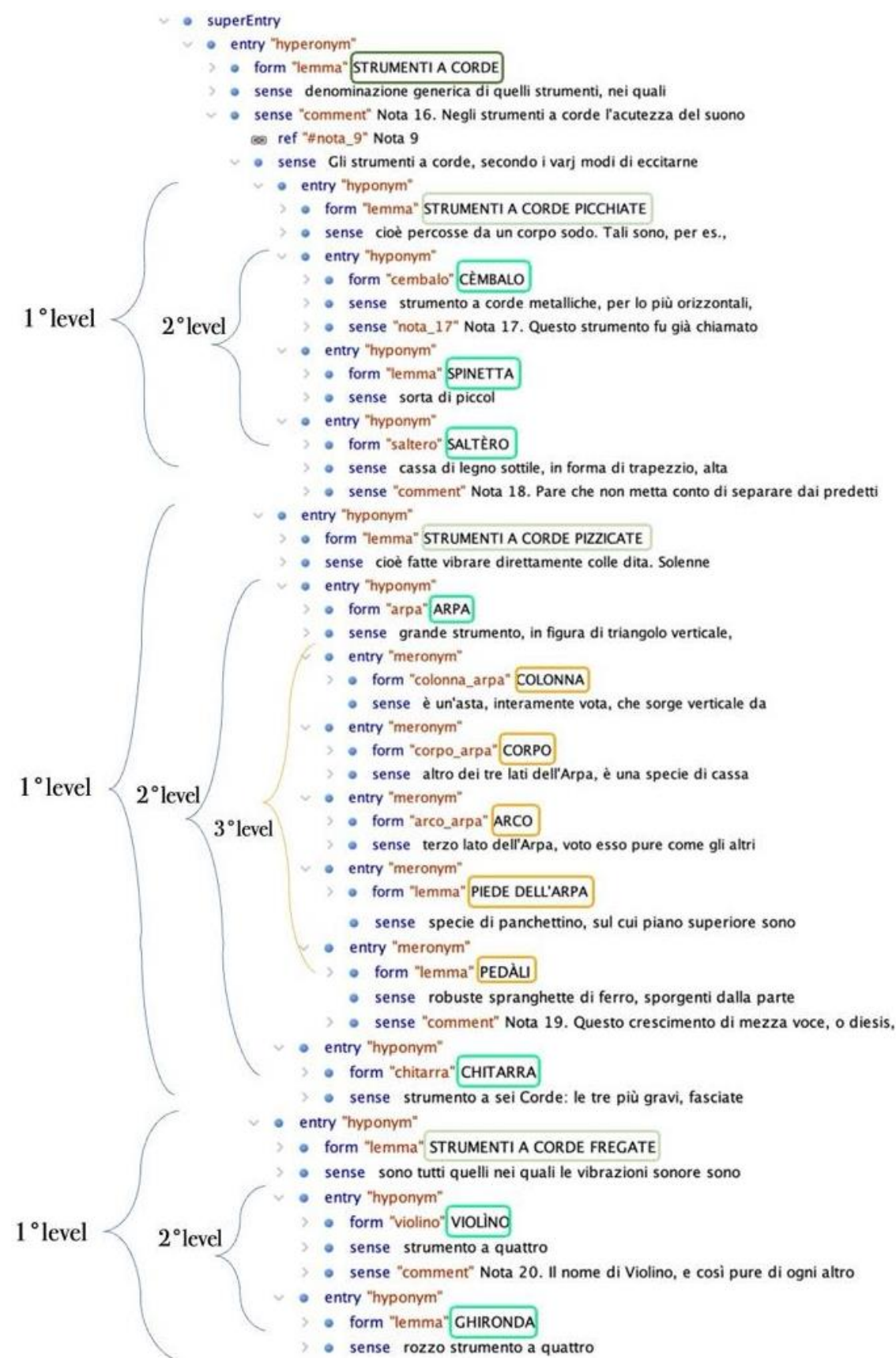


FIG. 2: *Oxygen XML Editor*'s outline showing markup of semantic relations and nesting levels (hypernyms-hyponyms; holonym-meronyms).



FIG. 3: The result of a search run on the experimental website hosting Italian non-alphabetic dictionaries in electronic format (enabled by XML-TEI markup): the lemma *vasellame da cucina*.

References:

- Biffi M. 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, Sintra, Portugal, 1-3 October 2019, Lexical Computing CZ, pp. 603–621.
- Biffi, M. (2024). *Per una terminologia condivisa dei dizionari elettronici/ digitali*. «LADINIA», vol. XLVIII, pp.53-70, ISSN:1124–1004.
- Khemakhem, M. et al. (2017). *Automatic Extraction of TEI Structures in Digitized Lexical Resources using Conditional Random Fields*, in *Electronic lexicography*, eLex 2017 (September 2017), Leiden, Netherlands.
- Khemakhem, M. et al. (2018). *Enhancing Usability for Automatically Structuring Digitised Dictionaries*, GLOBALEX Workshop at LREC 2018 (May 2018), Miyazaki, Japan.
- Patella, B. (2023). *Versione elettronica dei principali dizionari metodici ottocenteschi della lingua italiana*, tesi di Dottorato in “Filologia, Letteratura italiana, Linguistica” (curriculum: “Umanistica Digitale”), Università degli Studi di Firenze, a.a. 2022/2023.
- Patella, B. (2024a). *Codifica XML-TEI: proposta di mark-up per i dizionari metodici*, in AA., *I dati linguistici. Metodologie e strumenti della ricerca*, Studi e ricerche del Dipartimento di Lettere e Filosofia, Società Editrice Fiorentina, pp. 321-338. DOI: <https://doi.org/10.35948/DILEF/978-88-6032-750-5.17>
- Patella, B. (2024b). *La versione elettronica del Vocabolario toscano dell'arte del disegno (1681) di Filippo Baldinucci*, «Studi di Memofonte», vol. 34/2024, pp. 153-180. ISSN:2038-0488 https://www.memofonte.it/files/Studi-di-Memofonte/rivista30/XXXIII/XXXIII_2024_PATELLA.pdf
- Riccio, A. (2016). *Gli strumenti per la ricerca linguistica. Corpora, dizionari e database*, Roma,

Carocci

Salgado, A. et al. (2024). *The Morais Dictionary: Following Best Practices in a Retro-digitized Dictionary Project*, International Journal of Humanities and Arts Computing. 18(1), pp. 125–147. <https://doi.org/10.3366/ijhac.2024.0325>

TEI P5 (2025). *Guidelines for Electronic Text Encoding and Interchange*, [edited by] Text Encoding Initiative Consortium, gennaio 2025 (available on-line) <https://tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf>

Marta Petrak, Ivana Franić

Faculty of Humanities and Social Sciences, University of Zagreb, Croatia

mpetrak@ffzg.hr, ifranic1@ffzg.hr

FraCroVal: a contribution to contrastive studies of verbal valency

Keywords: verbs, verbs of perception, valency, corpus approach, contrastive linguistics

Verbal valency, one of the fundamental concepts of Tesnière's (1959) structural syntax, has become the basis for numerous studies of sentence syntax. In recent decades, valency theory has strongly influenced the linguistic description of natural languages. In this paper, we present a new contribution to the study of verbal valency from a contrastive point of view. In order to do so, we focus on a new French-Croatian valency database FraCroVal available as an online tool for researchers and language learners. Our basic structural framework was upgraded with basic tenets of Cognitive Linguistics and a general usage-based approach.

The database currently consists of French and Croatian verbs of visual perception, which were the first analysed group due to their major importance for the construal of the lexicon (Grezka 2009) and the overall structure of our conceptual system (Evans & Green 2006: 46). The verbs were retrieved from two large comparable web corpora, frWaC and hrWaC, and their meanings were analysed on the basis of their use in the context.

Our analysis has demonstrated the following: 1) French and Croatian verbs of visual perception exhibit very similar valency patterns, the most common one being the [[Experiencer] [Stimulus]] construction (Usoniene 1999: 211); 2) French verbs exhibit interesting diachronic dynamism on the semantic level; 3) the meanings of the majority of verbs are metaphorically mapped onto the domain of UNDERSTANDING (Sweetser 1990).

The findings were made publicly available through the FraCroVal database, which lists verbal meanings accompanied with examples of use in corpora and translations. Moreover, the database also serves as a hub providing various syntactic and semantic information from online sources. We hope that this innovative approach to the study of verbal valency might serve as an incentive for future bilingual valency research.

References:

- Evans, V. & Green, M. (2006) *Cognitive Linguistics. An Introduction*. Edinburgh University Press.
- Grezka, A. (2009) *La polysémie des verbes de perception visuelle*. L'Harmattan
- Sweetser, E. (1990) *From Etymology to Pragmatics. Metaphorical and Cultural Aspects of Semantic Structure*. Cambridge University Press.
- Tesnière, L. (1959) *Éléments de syntaxe structurale*. C. Klincksieck.
- Usoniene, A. (1999) Perception verbs revisited. *Working Papers*, 211–225.

Dóra Pődör

Károli Gáspár University of the Reformed Church in Hungary, Budapest

podor.dora@kre.hu

The First Irish-Hungarian / Hungarian-Irish Learner's Dictionary

Keywords: Irish-Hungarian Learner's dictionary, Hungarian-Irish Learner's dictionary, CEFR levels for Irish, Irish morphology for learners, geographical and style register labels for Irish words

The talk is going to present the editorial principles that have been employed to compile the first Irish-Hungarian / Hungarian-Irish Learner's Dictionary (levels A1-A2), which is due to be published at the end of 2025 or at the beginning of 2026. Some of the most important challenges will also be briefly mentioned, and sample entries will also be shown.

This is going to be the first bilingual dictionary for this language pair, and according to the author's knowledge, the first dictionary with more features that can facilitate the learning process than what are present in currently available Irish-English / English-Irish school dictionaries.

The aspects to be discussed are going to be the following: lemma selection; assigning A1 or A2 levels to headwords and/or meanings according to the Common European Framework of Reference for Languages (CEFR) (new feature); pronunciation information (partly new feature); the classification of verbs (improved); inflected forms (improved); information on initial mutations (these express grammatical relationships) (improved); geographical and style register labels (new feature); and grouping the headwords into various semantic and functional categories (similarly to what can be found in some monolingual English learner's dictionaries) (new feature).

References:

Beattie, S. (Ed.) (2016) *Collins Irish School Dictionary*. HarperCollins Publishers.

Ó Mianáin, P. (Chief ed.) (2020). *Concise English-Irish Dictionary*. Foras na Gaeilge.

Pődör, D. (2021) Pádraig Ó Mianáin (chief editor). 2020. *Concise English-Irish Dictionary Foclóir Béarla-Gaeilge. International Journal of Lexicography*, 35(1), 129–135. <https://doi.org/10.1093/ijl/ecab016>

Pődör, D. (2024) Egy ír–magyar/magyar–ír tanulószótár létrehozásának kihívásai. *Modern Nyelvoktatás* 2024/1–2., 155-167. <https://doi.org/10.51139/monye.2024.1-2.155.167>

TEG Levels. (2025) <https://www.teg.ie/teg-levels.8.html>

Rui Qiao
University of Montpellier Paul Valéry, France
qiaorui5201@gmail.com

Exploring Chengduhua through the Lexicographical Works of Missionaries (19th 20th Century)

Keywords: Historical lexicography, Chinese dialect studies, Corpus-based lexicography, Missionary linguistics, Chengduhua

Since the 17th century, missionaries in China have conducted linguistic research to facilitate communication and evangelization. Their efforts resulted in dictionaries, grammars, and pedagogical manuals, which document both the challenges and methodologies of learning Chinese. From the 19th century onwards, lexicographical works on various dialects appeared across China, offering valuable insights into dialect acquisition and description.

This study examines missionary-produced materials on Chengduhua (the dialect spoken in Chengdu), focusing on two key works: *Chinese Lesson for First Year Students in West China* (1917) by Omar L. Kilborn, a medical textbook for Western students, and *Western Mandarin or the Spoken Language of Western China* (1900) by Adam Grainger, a dictionary for newly arrived missionaries. Comparing these materials with modern Chengduhua reveals phonetic and lexical changes, including words that have evolved, disappeared, or been influenced by Mandarin and other dialects.

Kilborn's manual presents 1,002 practical phrases with English translations and Wade-Giles romanization, marking tones (1 to 4), while Grainger's dictionary features a five-tone notation system. These works demonstrate methodological adaptations in phonetic notation and dialect documentation.

This study contributes to historical lexicography by analyzing Chengduhua's phonetic and lexical features. It highlights the role of missionary lexicographical contributions in preserving and studying Chinese dialects, aligning with discussions on dialect dictionaries, corpus-based lexicography, terminology, and methodological innovations in language documentation.

References:

- Deng, Z. (2011). 传教士所编《西蜀方言》及其在四川方言研究中的价值 (Chuánjiàoshì suǒ biān Xīshǔ fāngyán jí qí zài Sìchuān fāngyán yánjiū zhōng de jiàzhí) [The missionary-compiled *Western Mandarin* and its value in Sichuan dialect studies]. *汉语史研究集刊* (Hànyǔ shǐ yánjiū jíkān) [Journal of Chinese Historical Linguistics], 233–244.
- Grainger, A. (1900) *Western Mandarin; or, The spoken language of Western China: With syllabic and English indexes*. Shanghai: American Presbyterian Mission Press.
- Huang, L. (2010) 再论钟秀芝《西蜀方言》的入声和基础音系问题 (Zàilùn Zhōng Xiùzhī Xīshǔ fāngyán de rùshēng hé jīchǔ yīnxì wèntí) [On the entering tone and the basic phonological system in Zhong Xiu-zhi's *Western Mandarin*]. *语言科学* (Yǔyán kēxué) [Linguistic Sciences], 9(4), 402–415.
- Kilborn, O. L. (1917) *Chinese lessons for first year students in West China*. [s.l.]: Union University.

You, R. (2002) 西洋传教士汉语方言学著作书目考述 (Xīyáng chuánjiàoshì Hànyǔ fāngyánxué zhùzuò shūmù kǎoshù) [*A study of bibliographies of Western missionaries' works on Chinese dialectology*]. 哈尔滨: 黑龙江教育出版社 (Hēilóngjiāng jiàoyù chūbǎnshè) [Heilongjiang Education Press].

Iasmin Valéria Miranda Rabelo¹, Adriana Silvina Pagano², Maucha Andrade Gamonal³

¹⁻²Universidade Federal de Minas Gerais, Brasil

³Universidade Federal de Juiz de Fora, Brasil

iasminvmr@ufmg.br, apagano@ufmg.br, maucha.andrade@visitante.ufjf.br

Disability and Assistive Technology: Modeling Accessibility Domain Frames on FrameNet Brasil

Keywords: Frame Semantics, FrameNet Brasil, Accessibility, Disabilities, Assistive Technology

This paper presents a lexical-semantic study of the Accessibility domain in Brazilian Portuguese, based on the theoretical foundations of Frame Semantics (FILLMORE, 1982) and the methodologies of the FrameNet project (RUPPENHOFFER ET AL, 2016). The aim of this research is to model frames from the Accessibility domain in FrameNet Brasil in order to develop a more comprehensive and inclusive database. To this end, this work highlights the importance of language as a tool for social inclusion through a history of studies on accessibility and connections established between Lexical Semantics and Terminology. The corpus of analysis was compiled using booklets, glossaries and other textual resources related to the area of accessibility. Through the analysis of this corpus, the main lexical items in this domain were identified, selected as candidate Lexical Units and organized into two frames: Deficiência (Disability) and Tecnologia_Assistiva (Assistive_Technology). The lexicographic annotation methods of FrameNet Brasil allowed a recognition of the main contexts of use of the LUs that evoke these frames, information that was used to determine the Frame Elements and also to create the third frame Pessoas_com_deficiência (People_with_disabilities). This research contributes to Accessibility studies within a linguistic context in Brazilian Portuguese. The development of a more comprehensive lexical resource, aware of issues such as ableism, prejudice and social integration, highlights the importance of language in maintaining a more inclusive society.

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

Institute of the Estonian Language, University of Tartu, Estonia

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

Who decides what's informal? Reducing subjectivity in dictionary labeling with corpora and LLMs

Keywords: large language models, corpus analysis, register labels, lexicography, Estonian

This presentation introduces the results of an experiment on how corpus analysis and large language models (LLMs) can assist lexicographers in assessing word informality for dictionary labeling. The study seeks to reduce subjective decisions in Estonian lexicography by analyzing large-scale linguistic data. The focus is on supporting the compilation of the *Dictionary of Standard Estonian* (DSE; forthcoming in December 2025; current edition: DSE, 2018). Labels indicating informality suggest that certain words or meanings are unsuitable for neutral and formal contexts. However, in the DSE, many words or meanings are labeled as informal for reasons other than actual informality (e.g., reducing linguistic variation, influence from other languages, etc.).

Before the wider use of corpora, register labels in Estonian dictionaries relied on individual intuition rather than on systematic analysis of actual language use. However, register analysis must not rely on the preferences of an individual but instead be based on samples that represent the register as comprehensively as possible (Biber & Conrad, 2009). Corpora and LLMs can support this process by providing a more objective basis, capturing language as used by the broader language community.

Our experiment tested 1,330 words from the forthcoming DSE to determine which words can genuinely be labeled as informal and to identify the reasons behind these classifications. The *Estonian National Corpus* (ENC, 2023; accessed via Sketch Engine) enables lexicographers to analyze word usage across genres representing different registers, and LLMs' (e.g., Claude 3.7 Sonnet, Gemini 2.5 Pro, GPT-4o) prompt can be directed to consider a word's occurrences in various types of contexts.

By integrating computational methods with human expertise, we aim to establish a more data-driven basis for the informality label. However, a key question remains: To what extent should a lexicographer's intuition influence labeling decisions? At present, lexicographers need to remain attentive to the limitations of computational tools.

References:

- Anthropic (2024) *Claude* [Large Language Model]
Anthropic. <https://www.anthropic.com/claude>
- Biber, D., & Conrad, S. (2009) *Register, Genre, and Style*. Cambridge University Press
- Estonian National Corpus (ENC) (2023) *Estonian National Corpus 2023*.
<https://doi.org/10.15155/3-00-0000-0000-0000-08C04M>
- Gemini Team, Google- (2024) *Gemini: A Family of Highly Capable Multimodal Models*. arXiv.
<https://arxiv.org/pdf/2312.11805>
- Raadik, M. (ed.), Erelt, T., Leemets, T., Mäearu, S., & Raadik, M. (comps.) (2018) *Dictionary of*

Standard Estonian 2018 (DSE 2018) EKSA.

Ewa Rodek

Institute of Polish Language Polish Academy of Sciences, Poland

ewa.rodek@ijppan.pl

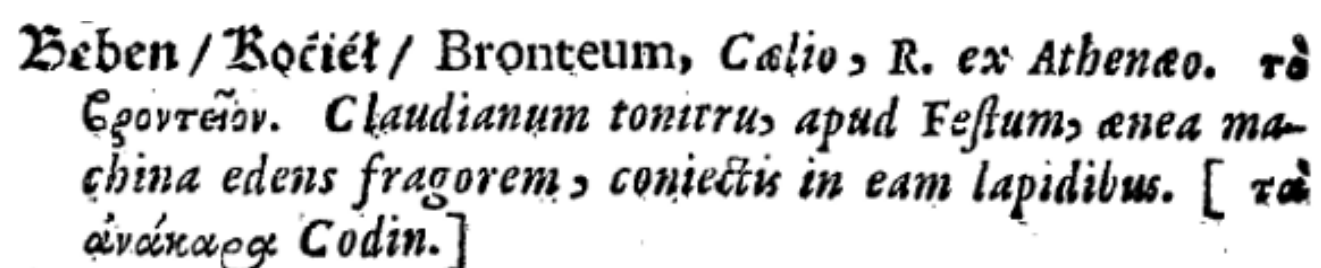
Training OCR Models for Historical Multilingual Dictionaries

Keywords: historical lexicography, TEI, digitization, double entry, OCR model

This presentation aims to demonstrate the process of training OCR models for multilingual historical dictionaries, as well as the challenges that must be addressed before such models can be trained.

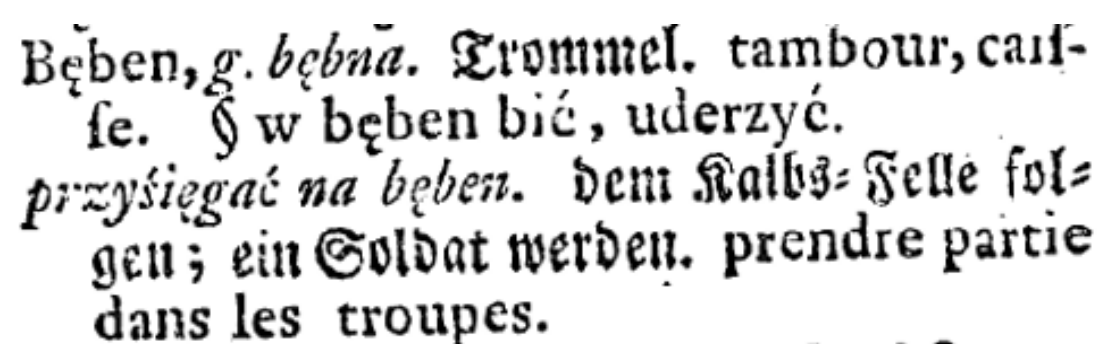
At the Institute of Polish Language Polish Academy of Sciences, we are implementing a project to create a Database of Historical Polish Lexicons. Every digitized dictionary will be encoded with TEI xml (in TEI Lex-0 standard) and presented in TEI Publisher as a search engine. It will be a tool for researchers to conduct multidirectional lexicographic research. We begin by digitizing three dictionaries: the Polish-Latin-Greek dictionary by Grzegorz Knapiusz (1643), the Polish-German-French dictionary by Michał A. Troc (1764) and the Polish-German dictionary by Jan Ernesti (1674).

These works are very extensive that were printed using various fonts and alphabets (figs. 1-3), and the available scans vary in quality. Therefore, creating models for OCR was challenging and required careful decisions. At the early stage of modeling, it was necessary, among other things, to address the authors' use of abbreviations and ensure their legibility for modern readers. In particular, Latin and Greek abbreviations required special treatment during model development. In addition, it was necessary to preserve as many punctuation marks as possible, whose proper configuration enables automatic TEI XML tagging. The dictionaries also contain double entries (fig. 4), which must be prepared for annotation at the modeling stage to ensure their content can later be displayed in a non-linear order. These and other challenges, along with their solutions, will be presented on the poster.



Beben / Bęciét / Bronteum, Calio, R. ex Athenæo. τὸ ἑγογγεῖον. Claudianum tonitru, apud Festum, ænea machina edens fragorem, coniectis in eam lapidibus. [τὰ αἰνάνια Codin.]

Fig. 1. Entry in the Polish-Latin-Greek dictionary by Grzegorz Knapiusz (1643)



Beben, g. bębna. Trommel. tambour, carafe. § w beben bić, uderzyć. przysięgać na beben. dem Kalb- Felle folgen; ein Soldat werden. prendre partie dans les troupes.

Fig. 2. Entry in the Polish-German-French dictionary by Michał A. Troc (1764)

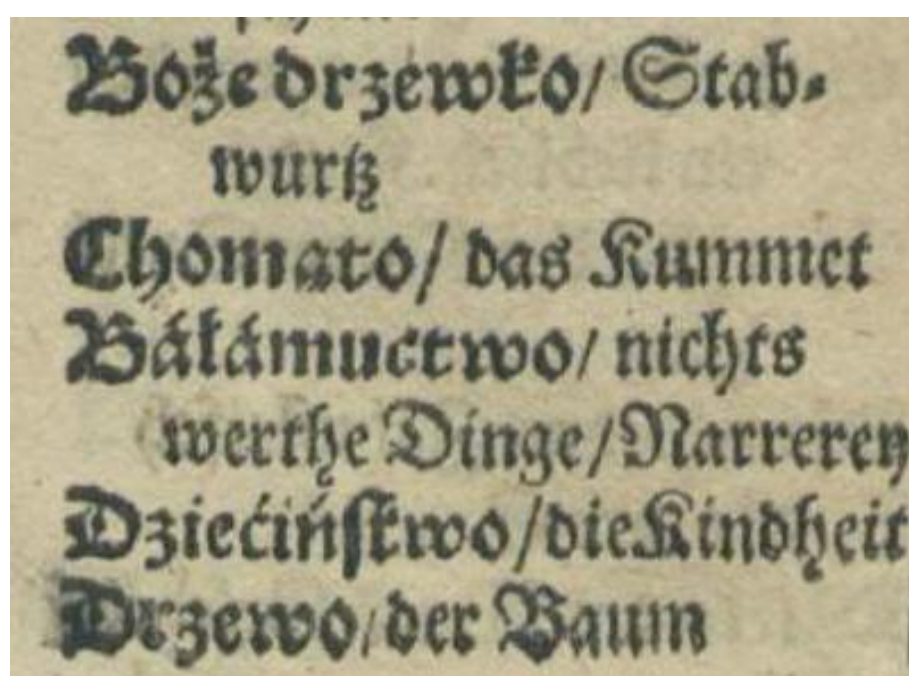


Fig. 3. Entry in the Polish-German dictionary by Jan Ernesti (1674)

Kradziefki. } 1) diebisch. 2) diebisch,
 Kradzieżny. } zum Stehlen geneigt. 1)
 de voleur, de larron. 2) enclin à dé-
 rober. § 1) kradzieżna ręka; kra-
 dzieška sztuka. 2) kradzieżnych się
 strzeż umysłów.

Fig. 4. Example of double entry in Troc's dictionary

The research was carried out as part of the project "Third-level digitization of large 17th- and 18th-century dictionaries: Creation of a database of historical Polish lexicons" (no. NPRH/DN/SP/0003/2023/12) financed by the Ministry of Science and Higher Education of Poland.

Manana Rusieshvili-Cartledge, Marine Makhatadze
Ivane Javakhishvili Tbilisi State University, Tbilisi, Georgia
manana.ruseishvili@tsu.ge, marine.makhatadze540@hum.tsu.edu.ge

**Exploring Sensory Lexis: Typological and Structural
Perspectives of the Georgian-English Thematic Dictionary**

Keywords: thematically-bound, lexis, metalexicography, cultural notes, semantic relations

The thematically arranged bilingual dictionary is still at a relatively early stage of its development. There remain ways in which the genre can be further refined. With technologies for information presentation that have been developed outside lexicography, e.g. search engines, hand-held information tools, etc., the range of possibilities for lexicography to present dictionary contents online has become much broader than it used to be in offline media. Lexicographers and web developers at Ivane Javakhishvili Tbilisi State University (TSU) and Arnold Chikobava Institute of Linguistics have embarked on creating the online resource – Georgian-English Thematic Dictionary, which combines elements from the bilingual, thematic, and pedagogical traditions in lexicography. The Georgian-English Thematic Dictionary is designed with practicality in mind, aiming to systematically compile and document thematically bound lexis (colour terms; weather; senses; position, movement and travel; food and drink, natural environment, etc.), establish English equivalents, provide illustrative examples and cultural notes, which frequently deal with the connotations and cultural associations of the lexemes. The method applied in the macrostructural coverage of the lemmata in the dictionary is based on metalexicographic considerations, which allowed us to explore the Georgian linguistic heritage. From the perspective of microstructural text constituents, we will demonstrate some samples of enriched entry structure on an example of the thematic category of senses and its sub-categories (touch, sight, hearing, smell and taste) in the Georgian-English Thematic dictionary. In this core of lexical information, semantic knowledge is shared and extended by semantic properties, including syntagmatic and paradigmatic thematic information. This concerns paradigmatic (synonyms, word families, stylistic markings) and syntagmatic information (collocations, idioms, usage notes, etc.). These features will be helpful in decoding, encoding and assimilative purposes and will enable its users to retrieve linguistically related categories of information coherently and problem-oriented.

References:

- Atkins, B. S., & Rundell, M. (2008) *The Oxford guide to practical lexicography*. Oxford University Press.
- Bergenholtz, H., & Nielsen, S. (2013) The treatment of culture-bound items in dictionaries. In *Dictionaries. An International Encyclopedia of Lexicography: Supplementary volume: Recent Developments with Special Focus on Computational Lexicography* (pp. 469–481). de Gruyter.
- Gouws, R. H., & Prinsloo, D. J. (2010) Principles and practice of South African lexicography. *African Sun Media*.
- Hartmann, R. R. (2013) *Mixed dictionary genres. Dictionaries. An International Encyclopedia of Lexicography*, 381–393.

- McCarthy, M. (1994) *Cambridge Word Routes Anglais-Français: Lexique thématique de l'anglais courant*. Cambridge University Press.
- Muller-Spitzer, C. (2013) *Textual structures in electronic dictionaries*. Gouws, Rufus H. et al.(eds.), 2013, 367–381.
- Stark, M. (2011) *Bilingual thematic dictionaries (Vol. 140)*. Walter de Gruyter.
- Tarp, S. (2013) Lexicographic functions. Tarp, S. (2013). *New developments in learner's dictionaries III: bilingual learner's dictionaries*.
- Wiegand, H. E., Gouws, R. H., Heid, U., & Schweickard, W. (2013) *Addressing and addressing structures in printed dictionaries*. Gouws, RH, U. Heid, W. Schweickard and HE Wiegand (Eds.), 2013, 273–314.
- Zgusta, L. (2003) Planning and organization of lexicographic work. *Lexicography: critical concepts*, 70-82. Zgusta, L. (2010) *Manual of lexicography (Vol. 39)*. Walter de Gruyter.

Ersilia Russo
University of Florence, Italy
ersilia.russo@unifi.it

Phraseology in the *Vocabolario del Fiorentino Contemporaneo* ('Vocabulary of contemporary Florentine')*

Keywords: Phraseology, Lexicography, Sociolinguistics, dialect lexicon, *Vocabolario del Fiorentino Contemporaneo*

The *Vocabolario del Fiorentino Contemporaneo* (VFC) project, initiated by the Accademia della Crusca in 1994, is currently being finalized. The transcriptions of the interviews conducted in the neighbourhoods of Santa Croce, San Frediano, and Rifredi are being analyzed to identify typical Florentine words and expressions. In fact, the vocabulary aims to document the lexical “specificity” of Florence by adopting a “differential” perspective, i.e. highlighting the absence (or partial presence) of overlap between Florentine and Italian lexicons, considering their historical interaction.

In the case of phraseology, despite the acknowledged “pan-Italian phraseological consonance” (Fanfani, 2018, p. 32), this separation becomes more evident, as phraseology provides a stronger reflection of social identity and serves as a repository of a community’s “view of the world” (Casadei, 1996, p. 29). If dialect can be seen as “an articulate procedure of definition and discovery of the sociolinguistic link between individual and community” (Binazzi, 2018, p. 205), then phraseology acts as the spokesperson for that procedure, reflecting customs, habits, and ways of thinking of those who use it and pass it down.

Starting from a broad understanding of the phraseological domain, which includes “syntagmas of varying extension, complexity, and structure” (Simone & Piunno, 2017: 15), this contribution seeks to illustrate the framework used to describe multi-word units in VFC. Special emphasis will be placed on proverbs and idioms, which — more than any other type of phraseological unit — “represent the strength, creativity, and vitality of the language over time” (Fanfani, 2024, p. 151). In light of the need to develop phraseographic tools dedicated to Romance diatopic varieties, this presentation aims to offer, through extensive exemplification, a method for the lexicographic treatment of phraseology, regarded as an integral component of the dialect lexicon.

*All quotations are translated into English by the author

References:

- Binazzi, N. (2018) Parole che diventano storia. L'esperienza del “Vocabolario del fiorentino contemporaneo” (VFC). In D’Onghia, L., Tomasin, L. (Eds.), *Etimologia e storia delle parole* (pp. 195–206). Franco Cesati.
- Casadei, F. (1996) *Metafore ed espressioni idiomatiche. Uno studio semantico sull’italiano*. Bulzoni.
- Fanfani, M. (2018) Fraseologia e dizionari. In Benucci, E., Capra, D., Vuelta García, S., Rondinelli, P. (Eds.), *Fraseologia, paremiologia e lessicografia* (pp. 27–55). Aracne.
- Fanfani, M. (2024) Sulla fraseografia Toscana. In Autelli, E., Konecny, C. (Eds.), *Fraseografia e metafraseografia delle varietà diatopiche. Studi in onore di Fiorenzo Toso* (pp. 145–

159.) *Linguistik online*, 125 (1) <https://doi.org/10.13092/lo.125.10790>
Simone, R., & Piunno, V. (2017) Combinazioni di parole che costituiscono entrata.
Rappresentazione lessicografica e aspetti lessicologici. *Studi e Saggi Linguistici* LV (2)
13–44.

John M. Ryan¹, Víctor Parra-Guinaldo²

¹University of Northern Colorado, USA, Colorado

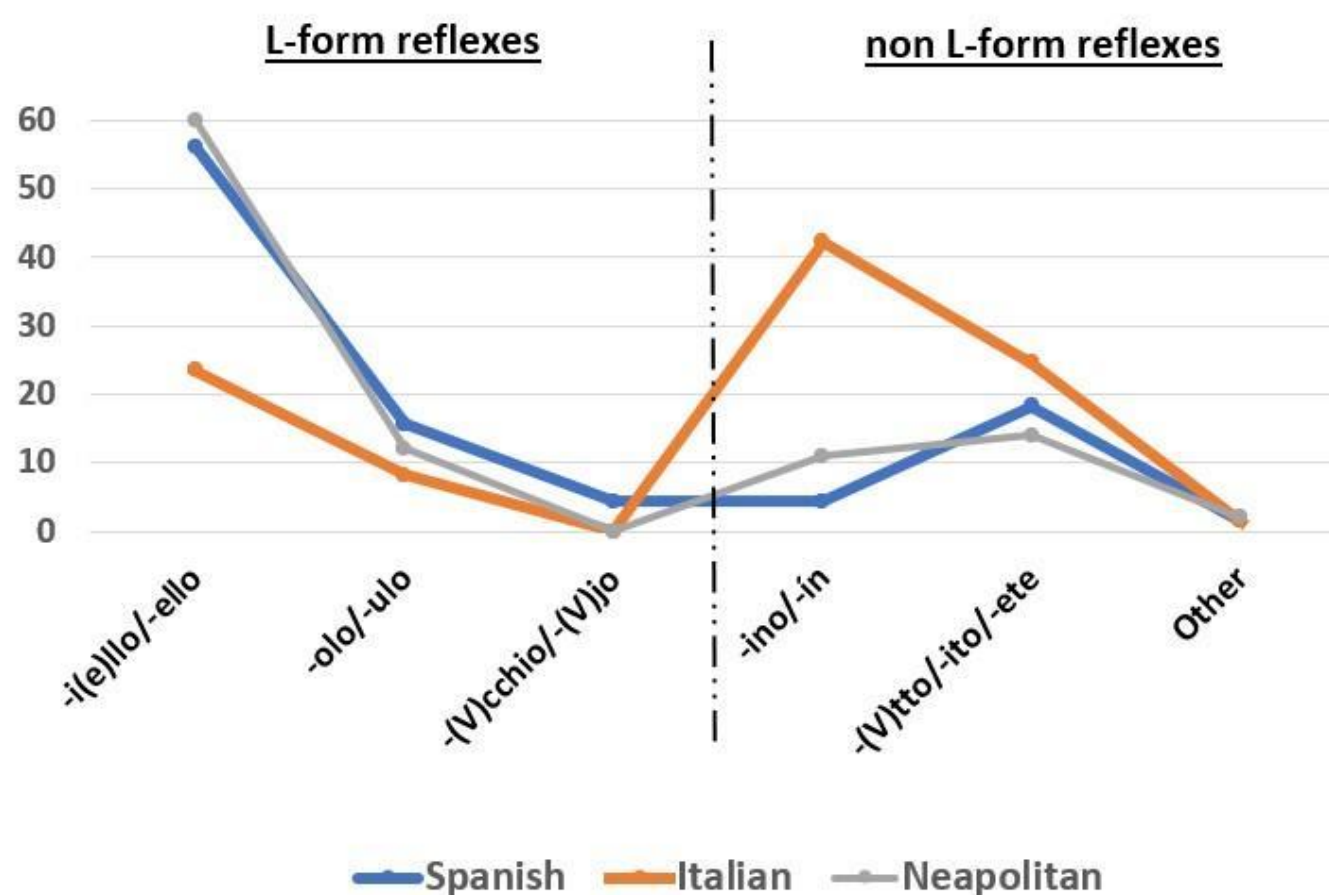
²Gulf University for Science & Technology, Kuwait

John.ryan@unco.edu, Parraguinaldo.V@gust.edu.kw

A Lexicographic Approach to the Classification of Relexified Diminutives in the Romance Languages: Phase III – Neapolitan

Keywords: dictionary studies, morphology, lexicon, diminutives, Neapolitan, Italian, Spanish

A distinctive morphological process that was quite productive in Latin and continues into the modern Romance languages and dialects is the employment of diminutive derivational suffixes to enhance the meaning of lexical word roots to which they are attached, particularly in informal and conversational speech. An additional historical feature of these root + suffix combinations over time is the loss of diminutive meaning of the suffix and ensuing semantic reanalysis of the combination as a new single root morpheme (e.g., Spanish *torta* ‘cake’ + *-illa* ‘little’ = *tortilla* not meaning ‘little cake’). This paper shows the results of Phase III of a larger project on diminutive relexification across the Romance languages by providing a quantitative lexicographic analysis of diminutives that have relexified in the history of the southern Italian dialect of Neapolitan. When compared to previous results for Spanish and Italian, namely, Phases I and II of the larger study, dictionary data suggest that Neapolitan has favored relexification with the *-(e)llo* suffix, in both Latin and modern periods, and although much like Italian and Spanish that have relexified with modern non-L-form reflexes such as *-ino/-ín* and *-etto/-ito*, it is unlike Italian in that Neapolitan has favored *-(e)llo* over *-ino*, making *-etto* slightly more common than *-ino*. The paper further supports the early Pan-Romance Diminutive Diasystem as asserted previously by the authors, suggesting that the same array of both L-form and non-L-form diminutive endings have served for purposes of diminutivization Romance-wide, but each language differs in accordance with the degree of contact between each region and the center of the Empire during the Latin era, as well as any ensuing contact among each other during the post-Latin period. Such was the four-hundred-year Spanish rule over the Kingdom of Naples, and the influence Spanish exerted on the Neapolitan lexicon during this period.



References:

- Ryan, J.M. & Parra-Guinaldo, V. (2016). Classification and history of relexified diminutives in modern Spanish: A lexicographic approach. In M.V. Rodríguez Domínguez et al. (eds), (2016). *Words across History. Advances in Historical Lexicography and Lexicology*. Las Palmas de Gran Canarias: Servicio de Publicaciones y Difusión Científica de la ULPGC, 364–380.
- Ryan, J.M. & Parra-Guinaldo, V. (2021). Spanish and Italian Diminutives Compared: Two Alternatives of a Single Diasystem. *Athens Journal of Philology* 8(1), 53–78.
- Ryan, J.M. & Parra-Guinaldo, V. (2023). Trends of Diminutive Relexification in Neapolitan: A Lexicographic Analysis with Comparisons to Spanish and Italian. *Athens Journal of Philology* 10(1), 9–34.

Bálint Sass

ELTE Research Centre for Linguistics, Research Institute for Lexicology, Hungary
sass.balint@nytud.elte.hu

Morphological dependency trees for representing constructions

Keywords: construction, constructicon, dependency, morphology

The Hungarian Constructicon (Sass, 2024) has a free text search functionality, i.e. the user is allowed to enter a short text in the user interface and the system extracts constructions from this text and shows their entries to the user. This extraction is based on automatic dependency parsing of both the input text and the constructions themselves present in the constructicon. The key step is matching the two dependency trees onto each other.

There are constructions in which not only syntactic but also some morphological features are relevant. For example, in *take part in something* the *part* element must be in the singular. If it happens to be in the plural, then it is not this construction. We also want to take these morphological features into consideration during matching.

According to the UD standard (de Marneffe et al., 2021) morphological features are represented as feature structures, i.e. a small table of feature-value pairs for every vertex of the tree. To create a unified data structure, we decided to transform the morphological information into a tree format. In the simple case, a one-level-deep tree fragment is created from the morphological features: features becoming edges and values becoming vertices, as seen in Fig. 1. Hierarchical morphological features can be represented straightforwardly by hierarchical tree fragments. In such a way we created a data structure which we call morphological dependency trees (MDTs).

MDTs are suitable to represent constructions including their syntactical and morphological properties both. Another big advantage of this homogeneous data structure is that any algorithm which works on trees can be applied to them without modifications, e.g. the matching algorithm mentioned above, improving its potentials in correctly identifying constructions.

References:

- de Marneffe, M. C., Manning, C. D., Nivre, J. & Zeman, D. (2021) Universal Dependencies. *Computational Linguistics* 47 (2) 255–308. <https://aclanthology.org/2021.cl-2.11/>
- 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, 275–283.

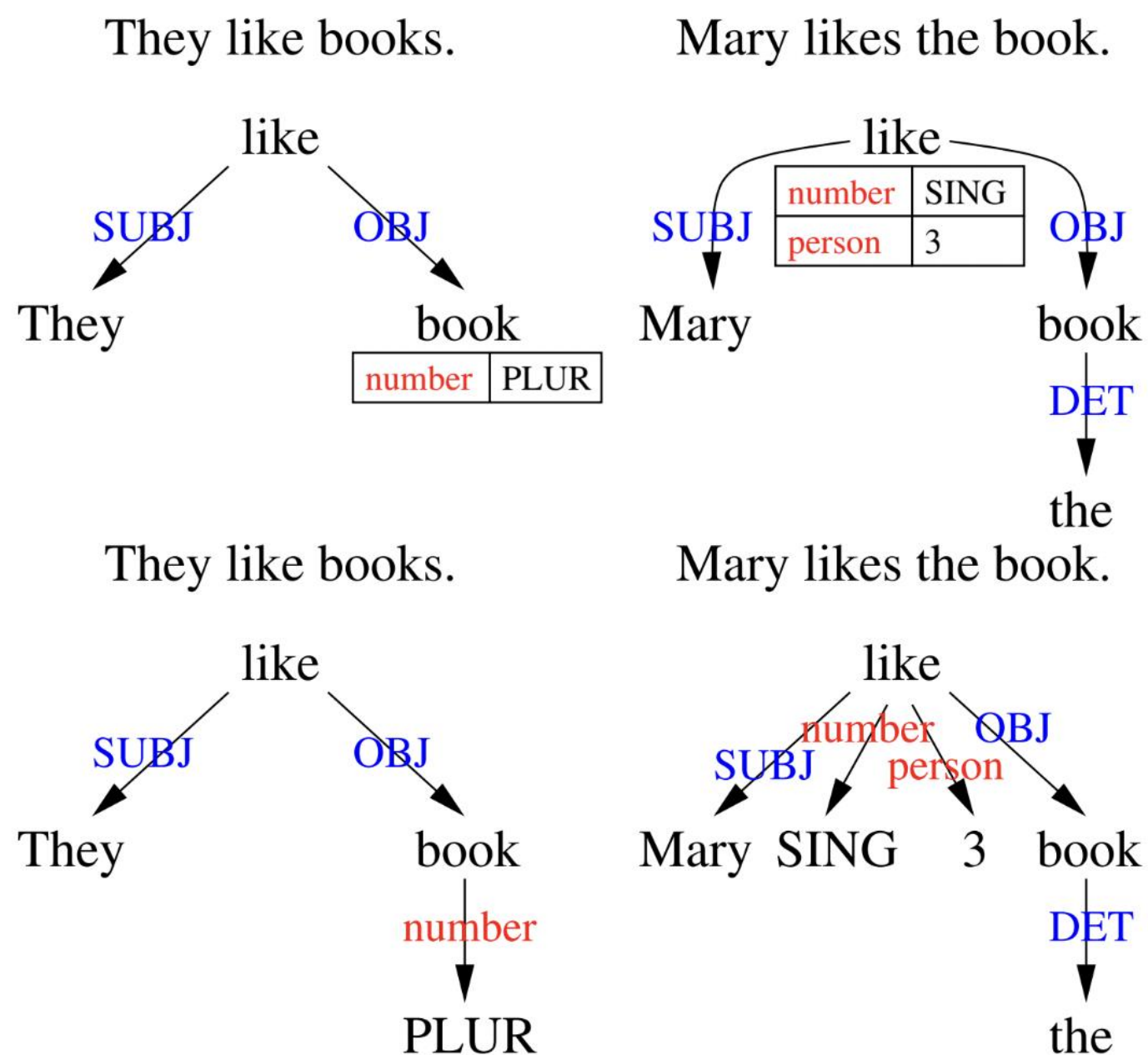


Figure 1: Traditional dependency trees (above) compared with morphological dependency trees (below). Syntactic information is shown in blue, morphological information is shown in red. It can be seen that all information is represented directly in the tree after the transformation. Examples are in English for anonymization purposes.

Olívia Seidl-Péché

Eötvös Loránd University Faculty of Humanities Institute of Language Mediation, Hungary
seidl-pech.olivia@btk.elte.hu

Terminological and Lexicological Challenges in Neural Machine Translation: A Case Study of EU Press Releases

Keywords: NMT, error, typology, terminological principles, correction

This presentation explores the theoretical implications of a recent terminological study by Hungarian scholars (the HITTL team: Robin et al., 2023) for the European Commission's Directorate-General for Translation (DGT). The research focuses on error patterns and correction strategies in post-edited and revised texts derived from neural machine translation (NMT).

Our corpus consists of ten EU press releases, each represented in three versions prepared by the DGT's Hungarian department:

- the raw NMT output produced by the EU's in-house *eTranslation* engine,
- the post-edited version by an institutional translator,
- the final version revised by a senior internal reviewer.

We systematically catalogued errors (Seidl-Péché & Kóbor, 2025) and human interventions using a dual annotation framework:

- error categories based on the DGT's internal typology (Müller, 2021) and the MQM core model (<https://themqm.org/the-mqm-typology/>),
- Robin's (2018) typology of human interventions.

Drawing on approximately thirty texts, we will illustrate:

- distinctive features of machine-generated errors,
- the challenges of distinguishing "terminological" errors from problematic "lexicological" choices (e.g. unidiomatic or awkward formulations).

The presentation will conclude with theoretical and practical insights on the interplay of terminology and lexicography in the context of machine translation, post-editing, and EU institutional translation practices.

References:

- Müller, R. (2021, March 10.) *Assessing the quality of translations: A practical guide*. [Conference presentation] Fifth ELRC Conference, online. <https://archive.lrc-coordination.eu/sites/default/files/2021-03/2.1%20-%20Assessing%20Quality%20in%20Translation.pdf>
- Robin, E. (2018) The classification of revisional modifications. In I. Horváth (Ed.), *Latest trends in Hungarian translation studies: Court interpreting, conference interpreting, terminology, audiovisual translation and revision* (pp. 155–163). Budapest: Hungarian Office for Translation and Attestation Ltd.
- Robin, E., et al. (2023, March 23.) *Human in the Translation Loop: az ELTE FTT és a DGT kutatási projektje*. [Conference presentation]. TransELTE 2023 Conference, Budapest, Hungary. <https://www.youtube.com/watch?v=3wPiqqvigjw>
- Seidl-Péché, O., & Kóbor, M. (2025) Terminológiai hibák és kategorizálásuk gépileg fordított és utószerkesztett szövegekben. In K. Fogarasi, D. Ittész, É. K. Varga, & T. Vágási

(Eds.), *Tudásmegosztás, információkezelés, alkalmazhatóság: II. Nyelvi közvetítés és beszéd kutatás* (pp. 108–116.) Budapest: Akadémiai Kiadó. <https://doi.org/10.1556/9789636640989.11>

Kganathi Shaku¹, Mmagonkahloleng Brudence Makau²

¹Department of Linguistics and Modern Languages, University of South Africa, Pretoria, South Africa

²Department of English and Comparative Literature, University of Fort Hare, East London, South Africa

shakukj@unisa.ac.za, mmakua@ufh.ac.za

Using digital platforms for data collection in lexicography: Mzansi Taal dictionary as case of analysis

Mzansi Taal dictionary is an online dictionary which records words and phrases from the colloquial languages used by the youth in South Africa. Just like other dictionaries, the lexicographic process of Mzansi Taal follows a robust step by step procedure for a word to end up in a dictionary. Traditionally, lexicographers would engage in reading of printed material like books and newspapers, reliance on self-experience and oral literature, to identify popularly used terms. Such process informs the lemmatization of words in the compilation process of a dictionary. Although technology brought more advanced digital tools such as social media, such tools are not fully exhausted and considered as effective methods of data collection in lexicography. It is without doubt that social media users use language, coin new words and phrases on daily basis as they are members of a digital linguistic community. As a result, social media streets have become a virtual village where language evolves. Therefore, if tools such as social media are used effectively, they have a potential of being instrumental for the purposes of data collection in lexicography. This makes it easy for modern lexicographers like Mzansi Taal team, to collect words and phrases effectively—with flexibility of time and cost effectively. The collection of words and phrases for Mzansi Taal dictionary demonstrated the effectiveness of using digital tools in the field of lexicography. Therefore, this paper explores the use of social media as an effective source of data collection in lexicography. The following colloquial languages are the research cases, Tsotsi Taal, Spitori, Ringas, Scamtho, Afrikaaps and Gayle. These colloquial languages are spoken in South African Township such as Soweto, Soshanguve, Mamelodi in Gauteng, District Six, Khayelisha and Cape Flats in the Western Cape. This is qualitative research, and it uses digital ethnography as the research inquiry.

References:

- Karelson, R. (1990) "Eesti kirjakeele seletussõnaraamat" tegija pilgu läbi. *Keel ja Kirjandus*, 1, 24–34.
- Kilgarriiff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014) The Sketch Engine: Ten years on. *Lexicography*, 1(1), 7–36. <https://doi.org/10.1007/s40607-014-0009-9>
- OpenAI. (2024) *GPT-4o system card*. arXiv. <https://arxiv.org/pdf/2410.21276>
- Risberg, L. (2024). *Sõnatähendused ja sõnaraamat. Kasutuspõhine sisend eesti keelekorraldusele* [The meanings of words and the dictionary. The impact of the usage-based approach on Estonian language planning] (Doctoral thesis) Tartu: University of Tartu Press.

Hindrik Sijens, Johan van der Zwaag
Fryske Akademy, Netherlands
hsijens@fryske-akademy.nl, jvdzwaag@fryske-akademy.nl

Towards a new historical dictionary of Frisian

Keywords: West Frisian, historical lexicography, electronic lexicography, dictionary revision, language periodization

The *Wurdboek fan de Fryske taal* (Dictionary of the Frisian Language, WFT) is a 25-volume lexicographical description of West Frisian (fry) between 1800 and 1975. According to the traditional periodization, this covers the New Frisian period. To fill in the lexicographical gap in the history of the Frisian language, *Fryske Akademy* is planning the compilation of a dictionary covering the Middle Frisian phase (1550-1800). However, this raises key questions and challenges which we will explore in this talk.

Traditionally, following other languages like Dutch, German and English, Frisian is divided into three language phases: Old Frisian (...-1550), Middle Frisian (1550-1800) and New Frisian (1800-...). This classification of Frisian is contested, because there are no major linguistic differences between Middle Frisian and New Frisian (Boersma 1999). Therefore, Middle Frisian is often referred to as Early New Frisian or Early Modern Frisian. At the same time, it is a crucial period for the development of West Frisian as a literary language and in shaping the contemporary Frisian culture. A lexicographical description of this language phase is therefore not only valuable in terms of continuity, but also important for research in linguistics and in the socio-cultural history of Friesland. A new dictionary that describes West Frisian from the period 1550-1800 creates challenges: Should a dictionary be created describing only this period or should Early Modern Frisian be integrated into the existing WFT? Or would it be better to opt for a completely new dictionary for the period 1550-1975? Should more recent material be included. Answering these questions has implications for the design of the dictionary: Should it be a 'traditional' dictionary or one that is conceptually based on modern, innovative insights?

References:

- Boersma, P. (1999) De leksikografy fan it Midfrysk. In A. Dykstra & R.H. Bremmer Jr (Eds.), *In skiednis fan 'e Fryske Taalkunde* (pp. 111-135). Fryske Akademy.
- Wurdboek fan de Fryske taal / Woordenboek der Friese taal. (1984-2011) Online edition: <https://gtb.ivdnt.org/search/>
- Fryske Akademy. *Frisian Corpora*. Retrieved August 26, 2025 from <https://frisian.eu/corpus-frontend/frysk/search/>

Sven-Erik Soosaar¹, Madis Jürviste², Tiina Paet³
^{1,2,3}Institute of the Estonian Language, Tallinn, Estonia
²University of Tartu, Tartu, Estonia
svenerik@eki.ee, madis.jurviste@eki.ee, tiina.paet@eki.ee

Large Language Models as Tools for Historical Lexicography: Automating Homonym Detection

Keywords: homonyms, historical lexicography, large language models, Estonian language

Our study contributes to the more extensive research topic on how language models (LLMs) could help automate dictionary compilation work (see also Lew 2023: 2; Jakubíček, Rundell 2023). Specifically, we examined polysemy and homonymy in one 17th-century Estonian dictionary, where spelling and meanings may differ from their modern interpretations.

The input for our experiment using Johannes Gutsclaff's German-Estonian dictionary (1648) consisted of words with identical forms or minor orthographic variations. We experimented using three LLMs: Gemini 2.0, Claude 3.5 Sonnet, and GPT-4o.

The experiment consisted of two parts. First, we had the model find modern Estonian literary language equivalents for Gutsclaff's 17th-century South Estonian words. In the second part, we had the model determine which of Gutsclaff's words with identical or similar forms were homonyms and which had the same meaning (in the reverse version of the dictionary, the same headword may appear multiple times).

As an illustrative example from the three LLMs tested, Claude performed most successfully, identifying homonyms even in cases where they are not homonyms in modern language. For instance, Gutsclaff's *selg* corresponds to modern Estonian *selg* 'back (body part)' and *selge* 'clear'. However, there were also errors, such as when the difference was in only one letter and sound. Despite identical spelling, the model failed to identify the pair for some word pairs. Claude identified most homonyms but also classified several word pairs as homonyms that should more accurately be considered polysemous words (e.g., *võõras* in the meanings of 'unknown' and 'guest'). Altogether 18 homonym pairs were identified correctly, and 5 pairs were not identified as homonyms by Claude 3.5 Sonnet. In addition, 38 pairs were identified as homonyms despite not being. Other models performed less accurately (corresponding numbers were 10-13-49 for GPT-4o and 12-14-84 for Gemini 2.0).

Our research demonstrates that large language models allow systematic identification of polysemous and homonymous entries in historical dictionaries. While LLMs also make mistakes in identifying homonyms and synonyms that a human would not make, these errors are nevertheless relatively easy to correct.

References:

- Gutsclaff, J. (1648) *Observationes grammaticae circa linguam esthonicam*. Dorpat: Johannes Vogel. <http://www.digar.ee/id/nlib-digar:100419> (10.09.2024)
- Jakubíček, M., Rundell, M. (2023) The End of Lexicography? Can ChatGPT Outperform Current Tools for Post-Editing Lexicography? *Proceedings of the eLex 2023 Conference: Electronic Lexicography in the 21st Century*. Brno: Lexical Computing, 508–523.
- Jürviste, M., Paet, T., Soosaar, S.-E. (2025) Eesti vanade sõnakujude tuvastamise võimalustest suurte keelemudelite abil. [Identifying Old Estonian Word Forms Using Large Language

Models.] *Eesti Rakenduslingvistika Ühingu aastaraamat 21. [Estonian Papers in Applied Linguistics 21*, 63–84.

Lew, R. (2023) ChatGPT as a COBUILD lexicographer. *Humanit Soc Sci Commun* 10, 704.
<https://doi.org/10.1057/s41599-023-02119-6>

Stefania Spina¹, Irene Fioravanti², Fabio Zanda³, Luciana Forti⁴, Damiano Perri⁵, Osvaldo Gervasi⁶

^{1, 2, 3, 4}University for Foreigners of Perugia, Italy

^{5, 6}University of Perugia, Italy

stefania.spina@unistrapg.it, irene.fioravanti@unistrapg.it, fabio.zanda@unistrapg.it
luciana.forti@unistrapg.it, damiano.perri@unipg.it, osvaldo.gervasi@unipg.it

A learner dictionary of Italian collocations: proficiency-based attribution and AI-generated definitions

Keywords: collocations, learner dictionary, L2 Italian, AI, proficiency label

This presentation describes the DICI-A, a learner dictionary of Italian collocations, which is aimed at filling a gap in the area of Italian lexicography, where none of the three existing collocation dictionaries (Lo Cascio 2013; Tiberii 2012; Urzì 2009) adopts corpus-based methods, nor is it specifically targeted at learners of L2 Italian. The DICI-A includes more than 10,000 collocations belonging to six syntactic relations: i. Verb + Direct object (*mantenere una promessa*, ‘to keep a promise’); ii. Adjective + Noun/Noun + Adjective, 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’).

Candidate collocations were extracted from PEC24 (Spina et al., 2025), a reference corpus of spoken and written Italian, combining pos-tagged and dependency parsed data. They were then filtered down by integrating statistical measures (frequency, dispersion, and association indices), and subsequently compared with existing dictionaries and assessed through human evaluation. The presentation will mainly focus on two lexicographically relevant features of the dictionary: i. as a dictionary targeted at learners, each collocation is assigned to a specific proficiency label, according to the Common European Framework of References for Languages (CEFR: Council of Europe 2020). This entails a range of choices involving the internal composition of collocations, their semantic transparency and their domain of use (García Salido et al. 2019); ii. definitions for the collocations included in the dictionary are obtained with the support of generative artificial intelligence systems. GenAI has proved effective in providing linguistically simple definitions, and thus potentially suitable for the dictionary’s target audience (Ptasznik et al. 2024). Methods used for these two specific features will be presented and discussed.

References

- Council of Europe (2020) Common European Framework of Reference for Languages: Learning, teaching, assessment – Companion volume, Council of Europe Publishing, Strasbourg, available at www.coe.int/lang-cefr
- García Salido, M., García, M., & Alonso-Ramos, M. (2019) Towards a Graded Dictionary of Spanish Collocations. In Electronic lexicography in the 21st century. *Proceedings of the eLex 2019 conference*. Brno: Lexical Computing, 849–864.
- 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), 322–336.
- Spina, S., Zanda, F., & Fioravanti, I. (2025) FROM PEC TO PEC24: A NEW REFERENCE CORPUS FOR ITALIAN. *Italiano LinguaDue*, 17(1), 745–768.
- Tiberii, P. (2012) *Dizionario delle collocazioni: Le combinazioni delle parole in italiano*, 2nd ed. Bologna: Zanichelli Editore.
- Urzi, F. (2009) *Dizionario delle Combinazioni Lessicali*. Lussemburgo: Convivium.

Mónika Varga

ELTE Research Centre for Linguistics, Institute for Historical and Uralic Linguistics, Hungary
varga.monika@nytud.elte.hu

From 'shape' to 'sort of' (from hedge to dress): a historical corpus study of *szabású*

Keywords: vague additive, corpus study, functional change, Middle and New Hungarian, dictionaries

This study investigates a formal and specific vague additive (Overstreet 2011/2012), *szabású* 'cut, sort of'. The earliest attested meaning of this element is 'shaped, -formed' in the 16th century. Meanwhile, it also became a hedge meaning 'sort of, kind of' and collocated with various targets including nouns, adverbs, and verbs (NySz., Simonyi 1918). During the 19th century, the meaning of *szabású* became more direct referring to the way a piece of clothing was cut (*bő szabású* 'loose fitted') or highlighting the similarity based on visual impressions (*emberszabású* 'anthropoid, man-like'). In standard Modern Hungarian, *szabású* has been lexicalized in expressions such as *nagyszabású* 'grand, large-scale', where it no longer functions as a hedge meaning 'sort of'. The aim of the present study is to unravel the semantic characteristics and plausible motivations for the hedge use of *szabású* between the 16th and 19th centuries. The research builds on data from historical databases (e.g. TMK, KED), including speech-related documents, which are rather understudied in this respect. More formal sources cited by the *Hungarian Historical Dictionary* (NySz.) regarding *szabású* were also taken into account. The analysis focuses on the collocational patterns of *szabású*, presuming that they varied across registers. In the disciplinary discourses of the 16th and 17th centuries, the typical targets were adjectives (especially those describing colours and other physical properties), while verbs occurred mostly in private letters and religious prose (e.g. *meguntam szabású* 'I got sort of bored with it'). The latter construction is specific to Hungarian, as other vague additives (e.g. *-szerű*, *-féle*, *-fajta* and *jellegű* 'kind of') do not tend to collocate or lexicalize with verbal targets. Data from the 18th and 19th centuries were also considered to describe the functional change of the element and how its non-lexicalized use became obsolete.

References:

- KED = Középmagyar emlékirat- és drámakorpusz [Corpus of Middle Hungarian Dramas and Memoirs] <https://ked.nytud.hu/#open>
- NySz. = Szarvas, G. – Simonyi, Zs. (1890–1893) *Magyar nyelvtörténeti szótár a legrégibb nyelvemlékektől a nyelvújításig 1–3* [Hungarian Historical Dictionary]. Hornyánszky Viktor.
- Overstreet, M. (2011/2012) Vagueness and hedging. In: Andersen, G. – Aimer, K. (eds.), *Pragmatics of society*. de Gruyter. 293–317.
- Simonyi, Zs. (1918) Bánom szabású és tarcabás. ['I am sort of sorry and tarcabás = sort of afraid'] *Magyar Nyelvőr* 47, 87.
- TMK= Történeti magánéleti korpusz [Old and Middle Hungarian Corpus of Informal Language Use]. <https://tmk.nytud.hu/>

Teresa Fuentes, Yuliia Vasik
University of Salamanca, Spain
tfuentes@usal.es, vasikyuliia@usal.es

Monolingual Lexicography and Visual Impairment: Parameters for More Accessible Dictionaries

Keywords: digital lexicography, visual impairment, screen readers, accessibility, universal design

Achieving a more inclusive society requires advancing universal design principles, which include Equitable Use, Flexibility in Use, Simple and Intuitive Use, Perceptible Information, Tolerance for Error, Low Physical Effort, and Size and Space for Approach and Use (<https://universaldesign.ie/about-universal-design/the-7-principles>). In digital lexicography, adherence to the Web Content Accessibility Guidelines (<https://www.w3.org/TR/WCAG21/>) is crucial, as web accessibility relies on four key principles: perceivability, operability, understandability, and robustness.

For individuals with visual impairments, screen readers are essential tools for web navigation. However, these assistive technologies often encounter limitations when accessing online dictionaries, resulting in incomplete or confusing lexical information. These issues primarily stem from three factors: technical constraints of screen readers, inadequate web accessibility standards tailored to the textual complexity of dictionaries, and lexicographic challenges related to the digital structuring of lexicographic data.

This study examines the microstructural components of dictionary entries in two prominent monolingual English dictionaries (<https://dictionary.cambridge.org/> and <https://www.merriam-webster.com/>), focusing on screen reader performance in two critical areas: (1) the interpretation of symbols and abbreviations, including pronunciation markers, and (2) the navigability of tabular structures distinguishing different word senses. By identifying the strengths and weaknesses of current dictionary configurations, this research aims to inform accessibility improvements and contribute to lexicographic practices that align with design-for-all principles.

Yelena Yerznkyan
Yerevan State University, Armenia
yerznkyan@ysu.am

Words with Pragmatic Loading: Main Principles of Dictionary Entry Development

Keywords: dictionary entry, word meaning, semantic information, pragmatic information, deixis

The present study takes up the issue of how pragmatic specifications encoded in linguistic signs as units of language system are provided in the lexicographic practice, as well as examines whether the way they are introduced defines them adequately and thus facilitates the speakers' ability to improve their pragmatic competence.

Assuming that the way dictionaries present the semantics of the word should take into account the type of meaning it conveys, the paper attempts to identify the semantic and pragmatic content, whether present or not, in the dictionary definitions of deictic words as one of the most vivid vocabulary layers with specific pragmatic loading.

The results obtained from an extensive study of monolingual English and Armenian dictionaries (including electronic) suggest that they show few signs of having directly addressed this issue. A notable discrepancy is observed in the overall number of descriptors (markers) of deicticity found in different dictionaries, which proves that this has not been done systematically.

As solving the identification problem in deictic reference involves primarily the setting of a basic reference point, the 'Origo' in Bühler's terminology, which points to the speech-situational factor, dictionaries can and should deal with it as the most essential and indispensable part of their meaning. The dictionary entry of a deictic term should signal the way the intended referent is to be decoded, it should organize that information according to the deictic categories with special reference to the type of the deictic sign and take into particular account the interaction between lexical semantics and pragmatics. Admitting that the reference point (the Origo) is the lexicalized pragmatic component, which is directly embedded in the meaning of deictic words and has a constant systemic status in language, it should be systematically integrated into the dictionary definitions as a marker of deicticity. The integration of this pragmatic information in the dictionary entries has become a necessity, as it totally governs the use of deictic expressions in speech.

References:

- Apresyan, Y. D. (1995) Pragmaticheskaya informatsiya dlya tolkovogo slovarya [Pragmatic Information for Monolingual Dictionaries]. In *Integralnoye opisanie yazika i sistemnaya leksikografiya [Integral Description of the Language and Systemic Lexicography]* (pp. 135–155). Shkola "Yaziki russkoy kulturi".
- Bühler, K. (1990) *Theory of language: The representational function of language*. John Benjamins.
- Yerznkyan, Y. L. (2013) *Deiksis slova: semantika i pragmatika [Word Deixis: Semantics and Pragmatics]*. YSU Publishing House.
- Zgusta, L. (1988) Pragmatics, lexicography and dictionaries of English. *World Englishes*, 7 (3) 243-253.

Carlo Zoli
Free University of Bozen-Bolzano, Italy
cazoli@unibz.it

Bridging Dialect Lexicography and Geolinguistic Atlases: the Smallcodes Approach

The *Smallcodes Platform*, recently awarded a prestigious recognition for the preservation of intangible cultural heritage (*Changes Awards, PNRR*), represents an advanced digital infrastructure for linguistic documentation, with a specific focus on diatopic (micro-)variation across phonetic, lexical, and syntactic dimensions. The platform addresses a critical gap in interoperability between heterogeneous methodologies employed in dialectology and areal lexicography. The lack of a standardized tool for both scholarly documentation and language activists engaged in language promotion constitutes a major challenge, particularly for less-documented and weakly standardized languages. Smallcodes is designed to function outside commercial constraints and missionary-driven approaches (*cf.* Dobrin, 2009), overcoming the technological and temporal limitations that typically hinder research initiatives and data collection efforts, whether within academic or crowd-sourced frameworks.

Traditionally, three methodologies underpin the documentation of diatopic micro-variation:

- Lexicographic documentation of individual local varieties;
- Comparative and areal lexicographic resources;
- Linguistic atlases (encompassing lexical, syntactic, phonetic, and ethnographic dimensions).

These methodologies rely on distinct software architectures, analytical frameworks, and dataformats, resulting in limited interoperability. The present study demonstrates how Smallcodes facilitates integration and convergence across these approaches by interfacing with key projects in Romance dialectology—such as AIS (AIS Reloaded; Loporcaro et al., 2021), ALI (a.o., Rivoira, 2017) the Manzini-Savoia Corpus (Mazzaggio et al., 2024; Savoia et al., 2025), and Alpilink (Rabanus et al., 2023) – as well as renowned lexicographic resources, including LSI (Lurà 2004), VLL (Videsott, 2020), and various Ladin dictionaries (Forni & Zoli, 2024).

By establishing an operational standard for data exchange between areal lexicographic resources and linguistic atlases, Smallcodes provides a scalable and future-proof solution for phonetic, lexical, and syntactic documentation. More than a theoretical model, a practical demonstration of the feasibility and transformative potential of interoperability in geolinguistic research, offering a replicable framework for cross-project data integration and collaborative linguistic inquiry.

References:

- Jaberg, K., & Jud, J. (1928–1940) *Sprach- und Sachatlas Italiens und der Südschweiz* (Vols. 1–8). Zofingen: Ringier u. C. <http://www3.pd.istc.cnr.it/navigais-web>
- Bartoli, M., et al. (1997–) *Atlante Linguistico Italiano* (Vols. published to date). Roma: Istituto Poligrafico e Zecca dello Stato. (Redaction at the University of Turin, supervised by L. Massobrio).
- Dobrin, L. M. (2009) SIL International and the disciplinary culture of linguistics: Introduction. *Language*, 85(3), 618–619. <https://doi.org/10.1353/lan.0.0157>
- Forni, M., & Zoli, C. (2024) L'integrazione tra grammatica e repertori lessicografici ladini. *Ladinia*, 48, 85–102.

- Kruijt, A., Rabanus, S., & Tagliani, M. (2023) The VinKo-Corpus: Oral data from Romance and Germanic local varieties of Northern Italy. In M. Kupietz & T. Schmidt (Eds.), *Neue Entwicklungen in der Korpuslandschaft der Germanistik: Beiträge zur IDS- Methodenmesse 2022* (Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache [CLIP], Vol. 11, pp. 203–212). Tübingen: Narr.
- Loporcaro, M., Schmid, S., Zanini, C., Pescarini, D., Donzelli, G., Negrinelli, S., & Tisato, G. (2021) AIS, reloaded: A digital dialect atlas of Italy and southern Switzerland. In A. Thibaut, M. Avanzi, N. Lo Vecchio, & A. Millour (Eds.), *Nouveaux regards sur la variation dialectale* (pp. 111–136). Strasbourg: Editions de Linguistique et de Philologie.
- Lurà, F. (2004) *Lessico dialettale della Svizzera italiana*. Bellinzona: Centro di dialettologia e di etnografia.
- Lurà, F., Moretti, M., & Zoli, C. (2009) Dalla carta al web: La versione informatica del lessico dialettale della Svizzera italiana. In G. Ruffino & M. D’Agostino (Eds.), *Storia della lingua italiana e dialettologia: Atti dell’VIII Convegno internazionale dell’ASLI*. Palermo: Centro di studi filologici e linguistici siciliani.
- Mazzaggio, G., Ludovico, L. A., Vena, M. V., Manzini, M. R., & Savoia, L. M. (2023) Morphosyntax of Italian and Romance varieties: Presentation of the Manzini and Savoia (2005) Corpus and its digitalization. *Bollettino dell’Atlante Linguistico Italiano*, 47, 185–210.
- Rabanus, S., Kruijt, A., Alber, B., Bidese, E., Gaeta, L., & Raimondi, G. (2023) *AlpiLinK Corpus 1.0.0* [Data set]. In collaboration with P. B. Mas, S. Bertollo, J. Casalicchio, R. Cioffi, P. Cordin, M. Cosentino, S. Dal Negro, A. Glück, J. Kokkelmans, A. Murelli, A. Padovan, A. Pons, M. Rivoira, M. Tagliani, C. Saracco, E. Siviero, A. Tomaselli, R. Videsott, A. Vietti, & B. Vogt.
- Rivoira, M. (2017) L’Atlante Linguistico Italiano. In *Il dialetto ad Umbertide* (pp. 16–19). Umbertide: Società Storica Umbertide Edizioni.
- Savoia, L. M., Manzini, M. R., Mazzaggio, G., Ludovico, L. A., Vena, M. V., Zoli, C., Baldi, B., Franco, L., & Binazzi, N. (2025) *The Manzini & Savoia (2005) Corpus: Morphosyntactic variation in Italian and Romansh dialects* [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.14803999>
- Videsott, P. (2020) *Vocabolar dl ladin leterar / Vocabolario del ladino letterario / Wörterbuch des literarischen Ladinischen*. Bozen-Bolzano: Bozen University Press. <https://doi.org/10.13124/9788860461681>
- Videsott, P., & Zoli, C. (2023) Il Corpus e il Vocabolar dl ladin leterar (Corpus e Vocabolario del ladino letterario / Corpus und Wörterbuch des literarischen Ladinisch). In *Perspectives de recherche en linguistique et philologie romanes* (pp. 1355–1366). Paris: Eliphi.

Krisztina-Mária Sárosi-Márdírosz¹, Krisztina Sófalvi²

¹Sapientia University of Transylvania, Faculty of Technical and Human Sciences, Department of Applied Linguistics, Târgu Mureş, Romania

²Sapientia University of Transylvania, Faculty of Technical and Human Sciences, Department of International Relations and European Studies, Cluj-Napoca, Romania

sarosikrisztina2015@gmail.com, krisztina.sofalvi@okt.kv.sapientia.ro

Hungarian-Romanian terminological matches in the Carpathian Basin educational terminology database

The terminological database subproject of the Termini Hungarian Language Research Network was launched in 2024 and focuses on the terminological standardization of the various educational terminology used in the countries neighbouring Hungary. The project is supported by the Hungarian Academy of Sciences in the framework of the Science for the Hungarian Language National Programme as a sub-programme entitled Hungarian Terminology Strategy, under the guidance of the HUN-REN Research Centre for Linguistics. The aim of the subproject is to create a database that collects and classifies Hungarian terms in various fields of education, taking into account the specificities of Hungarian language varieties from beyond the borders of Hungary. Another aim of the subproject is to create a database of educational terminology in nine languages (Hungarian, Croatian, German, Romanian, Serbian, Slovak, Slovenian, Ukrainian and English). As a member of the research team in Romania, we repeatedly face the challenges that arise from the differences between the educational systems in different countries. There are terms related to Hungarian education that have no Romanian equivalent, because the term, the institution, the structure, the method etc. does not exist in the Romanian educational system, or it exists in a different form. In this presentation I will show what solutions our research team uses to overcome these difficulties, in most cases by applying calques and giving definitions. I will also present some Romanian country- and region-specific terms.

Franck Zumstein

Université Paris Cité, ALTAE (URP 3967), Paris, France

franck.zumstein@u-paris.fr

A comparative study of authoritative 18th century pronunciation dictionaries of the English language

Keywords: English orthoepic dictionaries, 18th century, vowel/consonant pronunciations, lexical stress, variation

James Buchanan published two dictionaries in 1757 (Buchanan, 1757a; Buchanan, 1757b) and one in 1766 (Buchanan, 1766). In all three dictionaries, he paid much attention to indicating the stressed syllables of headwords and the phonetic values of vowels using diacritics such as macrons for tense vowels and breves for lax vowels. His last dictionary can be considered as the true ancestor of 20th and 21st century pronunciation dictionaries such as the famous *English Pronouncing Dictionary* (Jones, 2011). Some 25 years later, John Walker published *A Critical Pronouncing Dictionary and Expositor of the English Language* (Walker, 1791). In his dictionary, Walker commented on the pronunciations of many words via remarks included in the word entries. He regularly pointed to his predecessors' dictionaries to discuss some pronunciation issues at the time and Buchanan was one of his "authorities".

As part of an ongoing research project, Buchanan's as well as Walker's dictionaries have been digitized as TEI/XML files. It has thus been possible to retrieve 165 remarks in which Walker refers to one of Buchanan's dictionaries. When comparing all four dictionaries, we can conclude that Walker consulted either Buchanan, 1757a or Buchanan, 1757b. For example, when commenting on the phonetic value of the stressed vowel of the word <alienate>, Walker mentions that Buchanan "makes it short". Indeed, the vowel is marked as a lax vowel /ă/ in both dictionaries of 1757 whereas it is indicated as a tense vowel /ai/ in Buchanan, 1766, in line with Walker's transcription of the vowel.

With this presentation, we intend to analyse Walker's remarks in which he disagrees with Buchanan regarding the words' pronunciations. Doing so, we aim to assess whether such differences correspond to ongoing variations, or were just errors in Buchanan's 1757 dictionaries that were later corrected in his 1766 dictionary.

References:

- Jones, D. (2011) *English Pronouncing Dictionary*. 18th ed. (Peter Roach, James Hartman & John Easling, Eds.). Cambridge University Press. (Original work published 1917)
- Buchanan, J. (1757a) *Linguæ Britannicæ Vera Pronunciatio: or a New English Dictionary*. London: A. Millar.
- Buchanan, J. (1757b) *A New Pocket-Book for Young Gentlemen and Ladies, or a Spelling Dictionary of the English Language*. London: R. Baldwin, at the Rose, Paternoster Row.
- Buchanan, J. (1766) *Essay towards Establishing a Standard for an Elegant and Uniform Pronunciation of the English Language*. London: printed for E. & C. Dilly.
- Walker, J. (1791) *A Critical Pronouncing Dictionary and Expositor of the English Language*. London: G. G. J. and J. Robinson.