39th International Conference of the Croatian Association for Applied Linguistics

Pre-conference Day

Faculty of Humanities and Social Sciences, University of Zagreb
June 11, 2025

CLASSLA-express 2.0: Corpora vs. Large Language Models

Workshop on the Use of Corpora and Large Language Models in the Analysis of Language Constructions

Date and room: June 11, 9:00–13:00, Room A-102
Format: introductory talks and hands-on exercises
Number of participants: 30
Workshop convenors: Slobodan Beliga, Ivana Filipović Petrović, Jelena Parizoska

Details: The workshop will be given in Croatian language. Participation is free of charge. Please bring a laptop.

Registration: Interested participants are kindly asked to register by **June 3** via the form provided in the link.

Overview

In recent years, the number of user-friendly interactive tools drawing on large language models has grown rapidly. At the same time, linguistic research is increasingly examining the effectiveness of these tools in language-related tasks and their sensitivity to different types of prompts (De Schryver 2023; Rundell 2023; Jakubíček & Rundell 2023; Fuertes-Olivera 2024; Lew 2024; Davies 2025). Findings suggest that these technologies perform well in identifying and explaining the meanings of words and multi-word expressions, as well as in generating collocations. However, more demanding tasks, such as producing convincing examples and interpreting figurative language, remain a challenge, particularly in languages with limited resources (Filipović Petrović & Beliga 2024; Gantar 2024). In addition, concerns persist about the opaque nature of these systems, the unpredictability of their outputs, and their constantly evolving nature. In contrast, corpora continue to serve as a reliable and transparent source of linguistic data, allowing for the direct observation of authentic language use.

A key challenge for the development of digital linguistic infrastructure remains the underrepresentation of low-resource languages. As might be expected, large language technologies tend to be less effective for these languages, and corpus resources are often less developed. A major advance for South Slavic languages was the release of the CLASSLA web corpora in 2024 (Ljubešić & Kuzman 2024), supported by a series of workshops organised by CLARIN Slovenia across five South Slavic countries, aimed at training researchers in advanced corpus querying and analysis (Ljubešić et al. 2024).

This workshop builds on that initiative by adding a new dimension: bridging the benefits of corpora with the capabilities of tools powered by large language models. The focus will be on phraseme constructions, which combine fixed expressions with variable components and frequently carry figurative meaning. Participants will compare results retrieved from the CLASSLA corpora (using the NoSketch Engine concordancer) with those generated through interfaces that operate using large language models (such as ChatGPT, DeepSeek, Gemini, LeChat, among others). The workshop pursues two main objectives: to evaluate the effectiveness of language technologies in handling low- and medium-resource languages, and to explore their potential in processing complex multi-word expressions and semantic ambiguity. Ultimately, the goal is to contribute to a methodologically sound approach for integrating these two core tools, corpora and large language models, into contemporary linguistic research.

Agenda

Part I: Opening and Introductory Talks

09:00–09:10	Welcome
09:10–09:25	Nikola Ljubešić: *Web Corpora and Their Enrichment with Language Models*
09:25–09:40	Slobodan Beliga: *Large Language Models and Generative AI*
09:40–09:55	Ivana Filipović Petrović: *Applications of AI Tools in Phraseological Research*

Part II: Practical Session

09:55–10:40	Extraction of Language Data: Automatic Generation of Phrase Construction Lists Based on a Given Criterion
10:40–11:10	Coffee Break
11:10–12:00	Functional and Contextual Analysis of the Extracted Constructions
12:00–12:45	Processing and Classification of Language Data
12:45–13:00	Discussion

References

Beliga, Slobodan and Filipović Petrović, Ivana. 2024. Large Language Models Supporting Lexicography: Conceptual Organization of Croatian Idioms. In *Proceedings of the Conference on Language Technologies and Digital Humanities*, edited by Špela Arhar Holdt and Tomaž Erjavec, 23–46. Ljubljana: Institute of Contemporary History.

Davies, Mark. 2025. *Corpora and AI / LLMs: Overview*. https://www.english-corpora.org/ai-llms/corpora-vs-llms.html

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

Fuertes-Olivera, Pedro. 2024. Making Lexicography Sustainable: Using ChatGPT and Reusing Data for Lexicographic Purposes. *Lexikos,* 34, 123–140.

Gantar, Apolonija. 2024. Formulisanje rečničkih definicija pomoću veštačke inteligencije na primeru slovenačkih frazeoloških jedinica. In *Leksikografski susreti*, edited by Saša Marjanović, 151–158. Beograd: Filološki fakultet.

Jakubíček, Miloš, and Rundell, Michael. 2023. The End of Lexicography? Can ChatGPT Outperform Current Tools for Post-Editing Lexicography? In *Proceedings of the eLex 2023 Conference: Electronic Lexicography in the 21st Century*, edited by Marek Medveď et al., 508–523. Brno: Lexical Computing.

Lew, Robert. 2024. Dictionaries and Lexicography in the AI Era. *Humanities and Social Sciences Communications,* 11, 426.

Ljubešić, Nikola, and Kuzman, Taja. 2024. CLASSLA-web: Comparable Web Corpora of South Slavic Languages Enriched with Linguistic and Genre Annotation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024),* 3271–3282.

Ljubešić, Nikola, Kuzman, Taja, Filipović Petrović, Ivana, Parizoska, Jelena, and Osenova, Petya. 2024. CLASSLA-Express: A Train of CLARIN.SI Workshops on Language Resources and Tools with Easily Expanding Route. In *CLARIN Annual Conference Proceedings 2024,* edited by Vincent Vandeghinste and Thalassia Kontino, 31–35. Barcelona: CLARIN.

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