

ACL 2025

**10th Workshop on Slavic Natural Language Processing
(Slavic NLP 2025)**

**Co-located with the 63rd Annual Meeting of the Association
for Computational Linguistics (ACL)**

July 31, 2025

©2025 Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN 978-1-959429-57-9

Preface

This volume contains the papers presented at Slavic NLP 2025: the 10th Workshop on Natural Language Processing (NLP) for Slavic Languages. The workshop is organized by ACL SIGSLAV, the Special Interest Group of the Association for Computational Linguistics on NLP for Slavic Languages.

The Slavic NLP (formerly BSNLP) workshops have served as a key venue for over fifteen years, with a mission to advance the state of NLP for Slavic languages—languages that are spoken by more than 400 million people globally and represent an important part of the linguistic and cultural fabric of Eurasia.

The 2025 edition of the workshop continues the proud tradition established by the earlier BSNLP workshops, which were held in conjunction with the following venues:

- ACL 2007 Conference in Prague, Czech Republic.
- IIS 2009: Intelligent Information Systems, in Kraków, Poland.
- TSD 2011: 14th International Conference on Text, Speech and Dialogue in Plzeň, Czech Republic.
- ACL 2013 Conference in Sofia, Bulgaria.
- RANLP 2015 Conference in Hissar, Bulgaria.
- EACL 2017 Conference in Valencia, Spain.
- ACL 2019 Conference in Florence, Italy.
- EACL 2021 Conference in Kyiv, Ukraine.
- EACL 2023 Conference in Dubrovnik, Croatia.

Despite the importance and rich linguistic heritage of Slavic languages, the development of NLP tools and resources still lags behind those available for high-resource languages such as English. Many Slavic languages, particularly those spoken by smaller communities or lacking official EU status, remain underrepresented in both datasets and commercial NLP solutions.

Slavic languages present both practical and theoretical challenges: rich inflectional morphology, aspectual systems, and relatively free word order are among the many features that complicate parsing, translation, and generation. Yet, these features also create opportunities for research that is both linguistically informed and technically innovative.

Slavic NLP 2025 continues to serve as a forum for uniting researchers from academia and industry with a shared interest in advancing NLP for these languages. A recurring theme throughout this year's proceedings is the convergence of state-of-the-art methods—including large language models (LLMs), prompt-based learning, and multilingual transformers—with linguistically motivated problems, ranging from diachronic analysis to humor and persuasion detection.

This edition of the workshop features a diverse and ambitious research program. The accepted papers span a number of different Slavic languages and dialects, including Belarusian, Church Slavonic, Croatian, Czech, Macedonian, Polish, Russian, Slovak, Slovenian, and Rusyn. The topics address a rich spectrum of applications:

- Low-resource adaptation and language-specific modeling, including efforts to adapt definition modeling for Belarusian, build foundational models for Macedonian, and construct efficient summarization tools for Slovak.

- Diachronic and sociolinguistic analysis, through embeddings and curated datasets for Croatian news and Church Slavonic, as well as border effects in dialect variation for Rusyn.
- Bias and ethics in language technologies, such as examining gender representation in Czech and Slovenian LLM outputs and benchmarking LLM safety for Polish.
- Persuasion, propaganda, and disinformation detection, a central theme in this year’s Shared Task and related contributions, with a variety of generative, multitask, and ensemble approaches.
- Speech and syntax-focused work, including filled pause detection across South and West Slavic languages and high-efficiency transformer-based speech models.

A highlight of Slavic NLP 2025 is the Shared Task on the Detection and Classification of Persuasion Techniques in Parliamentary Debates and Social Media, attracting a record number of participants. This task aligns directly with the global relevance of NLP for combating manipulation and fostering informed public discourse. Several papers in this volume present novel methods –ranging from multitask debiasing to explanation-based data augmentation– for tackling this complex challenge.

Altogether, this volume features papers selected through rigorous peer review. The contributions represent a mix of theoretical insights, resource development, and practical systems, offering a snapshot of current research and emerging directions in Slavic NLP.

We thank all authors for their excellent submissions, the reviewers for their thoughtful and constructive feedback, and the participants of the Shared Task for their enthusiasm and innovation. We also express our gratitude to the organizing committee and program chairs for their commitment to ensuring the continued success of Slavic NLP.

We hope that this collection will be a valuable resource for researchers and developers, and that it will continue to foster collaboration across the many communities interested in the Slavic languages and their computational study.

The SlavNLP Organizers: Jakub Piskorski, Preslav Nakov, Nikola Ljubešić, Pavel Přibáň, Roman Yangarber

Program Committee

Chairs

Nikola Ljubešić, Jožef Stefan Institute
Michal Marcinczuk, Samurai Labs
Preslav Nakov, Mohamed bin Zayed University of Artificial Intelligence
Jakub Piskorski, Polish Academy of Sciences
Pavel Přibáň, University of West Bohemia, Faculty of Applied Sciences
Roman Yangarber, University of Helsinki

Program Committee

Zeljko Agic, Unity Technologies
Ekaterina Artemova, Toloka.AI
Dimitar Dimitrov, University of Sofia St. Kliment Ohridski"
Filip Dobranić, Institute of Contemporary History
Marina Ernst, University of Koblenz
Radovan Garabik, L. Stur Institute of Linguistics, Slovak Academy of Sciences
Jacek Haneczok, Erste Group IT
Milos Jakubicek, Lexical Computing
Mikhail Kopotev, University of Helsinki
Ivan Koychev, Sofia University St. Kliment Ohridski"
Vladislav Kubon, Charles University
Gaurav Kumar, University of California San Diego
Wojciech Kusa, NASK National Research Institute
Arkadiusz Modzelewski, Polish-Japanese Academy of Information Technology
Ivo Moravski, Sofia University
Maciej Ogrodniczuk, Institute of Computer Science, Polish Academy of Sciences
Petya Osenova, Sofia University St. Kl. Ohridskiand IICT-BAS
Alexander Panchenko, Skolkovo Institue of Science and Technology
Lidia Pivovarova, University of Helsinki
Senja Pollak, Jožef Stefan Institute
Marko Robnik-Sikonja, University of Ljubljana, Faculty of Computer and Information Science
Alexandr Rosen, Charles University, Prague
Agata Savary, Paris-Saclay University
Serge Sharoff, University of Leeds
Marko Tadić, University of Zagreb, Faculty of Humanities and Social Sciences
Marcin Woliński, Institute of Computer Science, Polish Academy of Sciences
Daniel Zeman, Charles University, Faculty of Mathematics and Physics

Table of Contents

<i>Identifying Filled Pauses in Speech Across South and West Slavic Languages</i> Nikola Ljubešić, Ivan Porupski, Peter Rupnik and Taja Kuzman	1
<i>Few-Shot Prompting, Full-Scale Confusion: Evaluating Large Language Models for Humor Detection in Croatian Tweets</i> Petra Bago and Nikola Bakarić	9
<i>GigaEmbeddings — Efficient Russian Language Embedding Model</i> Egor Kolodin and Anastasia Ianina	17
<i>PL-Guard: Benchmarking Language Model Safety for Polish</i> Aleksandra Krasnodebska, Karolina Seweryn, Szymon Łukasik and Wojciech Kusa	25
<i>Dialects, Topic Models, and Border Effects: The Rusyn Case</i> Achim Rabus and Yves Scherrer	38
<i>Towards Open Foundation Language Model and Corpus for Macedonian: A Low-Resource Language</i> Stefan Krsteski, Borjan Sazdov, Matea Tashkovska, Branislav Gerazov and Hristijan Gjoreski	44
<i>Towards compact and efficient Slovak summarization models</i> Sebastian Petrik and Giang Nguyen	58
<i>Adapting Definition Modeling for New Languages: A Case Study on Belarusian</i> Daniela Kazakouskaya, Timothee Mickus and Janine Siewert	69
<i>Bridging the Gap with RedSQL: A Russian Text-to-SQL Benchmark for Domain-Specific Applications</i> Irina Brodskaya, Elena Tutubalina and Oleg Somov	76
<i>Can information theory unravel the subtext in a Chekhovian short story?</i> J. Nathanael Philipp, Olav Mueller-Reichau, Matthias Irmer, Michael Richter and Max Kölbl	84
<i>When the Dictionary Strikes Back: A Case Study on Slovak Migration Location Term Extraction and NER via Rule-Based vs. LLM Methods</i> Miroslav Blšták, Jaroslav Kopčan, Marek Suppa, Samuel Havran, Andrej Findor, Martin Takac and Marian Simko	91
<i>DIACU: A dataset for the DIACHronic analysis of Church Slavonic</i> Maria Cassese, Giovanni Puccetti, Marianna Napolitano and Andrea Esuli	101
<i>Characterizing Linguistic Shifts in Croatian News via Diachronic Word Embeddings</i> David Dukić, Ana Barić, Marko Čuljak, Josip Jukić and Martin Tutek	108
<i>What Makes You CLIC: Detection of Croatian Clickbait Headlines</i> Marija Andelic, Dominik Sipek, Laura Majer and Jan Snajder	116
<i>Gender Representation Bias Analysis in LLM-Generated Czech and Slovenian Texts</i> Erik Derner and Kristina Batistič	124
<i>REPA: Russian Error Types Annotation for Evaluating Text Generation and Judgment Capabilities</i> Alexander Pugachev, Alena Fenogenova, Vladislav Mikhailov and Ekaterina Artemova	136
<i>Fine-Tuned Transformers for Detection and Classification of Persuasion Techniques in Slavic Languages</i> Ekaterina Loginova	151

<i>Rubic2: Ensemble Model for Russian Lemmatization</i>	
Ilia Afanasev, Anna Glazkova, Olga Lyashevskaya, Dmitry Morozov, Ivan Smal and Natalia Vlasova	157
<i>Gradient Flush at Slavic NLP 2025 Task: Leveraging Slavic BERT and Translation for Persuasion Techniques Classification</i>	
Sergey Senichev, Aleksandr Boriskin, Nikita Krayko and Daria Galimzianova	171
<i>Empowering Persuasion Detection in Slavic Texts through Two-Stage Generative Reasoning</i>	
Xin Zou, Chuhan Wang, Dailin Li, Yanan Wang, Jian Wang and Hongfei Lin	177
<i>Hierarchical Classification of Propaganda Techniques in Slavic Texts in Hyperbolic Space</i>	
Christopher Brückner and Pavel Pecina	183
<i>Team INSAntive at SlavicNLP-2025 Shared Task: Data Augmentation and Enhancement via Explanations for Persuasion Technique Classification</i>	
Yutong Wang, Diana Nurbakova and Sylvie Calabretto	190
<i>LLMs for Detection and Classification of Persuasion Techniques in Slavic Parliamentary Debates and Social Media Texts</i>	
Julia Jose and Rachel Greenstadt	202
<i>Fine-Tuned Transformer-Based Weighted Soft Voting Ensemble for Persuasion Technique Classification in Slavic Languages</i>	
Mahshar Yahan, Sakib Sarker and Mohammad Islam	217
<i>Robust Detection of Persuasion Techniques in Slavic Languages via Multitask Debiasing and Walking Embeddings</i>	
Ewelina Ksiezniak, Krzysztof Wecel and Marcin Sawinski	224
<i>Multilabel Classification of Persuasion Techniques with self-improving LLM agent: SlavicNLP 2025 Shared Task</i>	
Marcin Sawinski, Krzysztof Wecel and Ewelina Ksiezniak	231
<i>SlavicNLP 2025 Shared Task: Detection and Classification of Persuasion Techniques in Parliamentary Debates and Social Media</i>	
Jakub Piskorski, Dimitar Dimitrov, Filip Dobranić, Marina Ernst, Jacek Haneczok, Ivan Koychev, Nikola Ljubešić, Michal Marcinczuk, Arkadiusz Modzelewski, Ivo Moravski and Roman Yangerber	254