Брой 1 (21), Година 11 (2025)
Issue 1 (21), Year 11 (2025)

„DIGITALE SLAWISTIK“ PANEL AT THE 48TH AUSTRIAN LINGUISTICS CONFERENCE (ÖLT) 2024

Anna Jouravel
Albert-Ludwigs-Universität Freiburg (Germany)
ORCID ID: 0000-0002-7767-4973
anna.jouravel@slavistik.uni-freiburg.de

The panel "Digitale Slawistik" (Digital Slavic Studies) was held as part of the 48th Austrian Linguistics Conference (Österreichische Linguistiktagung) on December 18-19, 2024, bringing together scholars at the intersection of Slavic linguistics and digital humanities. The two-day event showcased a diverse range of digital approaches to various topics in Slavic language studies, from corpus-based analyses of contemporary social phenomena to the development of computational tools for processing historical Slavic texts.

The panel opened with a brief welcome from the organizers **Prof. Dr. Elias Bounatirou** (University of Vienna), **Dr. Anna Jouravel** (University of Freiburg), **Maximilian Grübsch** (University of Vienna) and **Ilia Afanasev** (University of Vienna), setting the stage for two days of scholarly exchange focused on digital approaches to Slavic linguistics.

In their introduction, the organizers emphasized that the implementation of digital technologies remains a significant desideratum in the field of Slavic studies, especially concerning language history and premodern Slavic written heritage, despite these technologies no longer being novel in other humanities disciplines. The panel specifically aimed to bring together scholars who are already using or developing methods and digital tools for Slavic linguistic research to address substantive research questions that can be answered using digital methods. An important goal of the panel was to move Slavic studies beyond the perpetual methodological discussions and debates about advantages and disadvantages, and instead focus on applying the now available robust tools for data collection, analysis, and evaluation—tools that have become indispensable parts of modern research—to scholar’s own research and specific linguistic questions.

The following contributions were presented at the panel in chronological order:

Regina Guzaerova, a postdoctoral researcher at the Institute of Slavic Studies at the University of Gießen (Germany) specializing in sociolinguistics, presented findings from her comprehensive corpus-based analysis of political correctness and new ethics concepts in Russian media. Her research employed advanced NLP techniques alongside traditional corpus linguistic methods to examine how these concepts have been represented and evolved in Russian media from 2010 to 2024.

Guzaerova's study utilized a diverse corpus compiled from Russian newspapers, online news platforms, blogs, and social media. Her methodological approach combined sentiment analysis to assess public attitudes, Named Entity Recognition (NER) to identify key entities, and Topic Modeling to uncover underlying themes. Through quantitative discourse analysis, she empirically demonstrated how media framing of these concepts varies by political orientation and media type.

The team, represented by **Maksim Aparovich**, PhD candidate in the Knowledge Technology Research Group (KNOT) at Brno University of Technology (Czech Republic), presented their work on developing a GLUE-type benchmark for Belarusian, filling a gap in resources for this East Slavic language. The researchers introduced five novel datasets focusing on sentence-level sentiment analysis, NER, linguistic acceptability, word-in-context assessment, and the Winograd schema challenge.

For sentiment analysis, they manually selected positive and negative sentences from diverse online sources; for NER, they worked with the be_hse corpus in Universal Dependencies (UD); for linguistic acceptability, they harvested unacceptable sentences from various sources including textbooks and machine translation outputs; for word-in-context assessments, they derived contexts from a Belarusian explanatory dictionary; and for the Winograd schema challenge (WSC), they expertly translated the original WSC-285 into Belarusian.

Preliminary evaluations showed no significant performance gap between commercial offerings like GPT-4o and free LLMs like Llama3-70B. Interestingly, their findings suggested that while sentiment analysis and word-in-context tasks appear to be mostly solved for Belarusian, models still struggle with acceptability judgments, NER, and especially the WSC.

Natalia Cheilytko, a postdoctoral researcher at the Institute of Slavistics and Caucasus Studies at the University of Jena (Germany), presented research examining Ukrainian in the 20th century as a multistandard language, following Peter Auer's theoretical framework. Their collaborative study highlighted how Ukrainian was characterized by two distinct standard varieties in the early 20th century, resulting from its division between Poland and the Russian Empire, and how these varieties began to converge following World War II despite the persistence of regional linguistic features, particularly in western and central Ukraine.
Cheilytko described their project using lectometry to investigate the mechanisms and trends underlying this convergence process through a large, representative linguistic corpus. She explained their methodology of calculating aggregated uniformity to measure diachronic distances among language varieties, focusing on onomasiological profiles of approximately 100 concepts that reflect the distribution of synonymous lexemes across different lects. For polysemous lexemes, they considered only occurrences where the lexeme was used in the relevant sense representing the target concept, with word sense disambiguation conducted using the GPT-4o-mini model.

The presentation detailed how their regionally annotated data was sourced from the GRAC corpus, and how they visualized the trajectory of convergence over time by analyzing pairwise distances among lects using splits graphs.

Liudmila Radchankava, a postdoctoral researcher at the Institute of Slavic Studies at the University of Mainz (Germany) presented research on complex adpositions in Russian, explaining how digital linguistic approaches can help analyze the development and structural dynamics of these linguistic elements. Her methodological approach included the use of extensive corpus data from the Russian National Corpus and the application of NLP techniques, particularly BERT, to form semantic clusters that capture different usage patterns and contextualizations of complex adpositions.

Her presentation focused on diachronic development and semantic changes of secondary adpositions through frequency analyses that documented changes in frequency and usage across different time periods. By combining digital analysis techniques with linguistic methods, Radchankava examined patterns of adposition formation and their syntactic functionality, contributing to a better understanding of their role within the Russian language system.

Edyta Jurkiewicz-Rohrbacher is currently Assistant Professor (Juniorprofessor) at the Institute of Slavic Studies, Universität Hamburg (Germany). From a user perspective, her presentation addressed how researchers can access the language competence of pre-trained

generative models for linguistic research. She postulated that translation should be considered a reliable task for indirectly accessing the linguistic competence of these models, similar to psycholinguistic tests and translational questionnaires used in theoretical and typological studies.

Using Russian bipredicative structures with adjacent personal pronouns in the dative case as a test case, she demonstrated how these ambiguous structures could reveal the syntactic parsing abilities of language models. Her study evaluated seven commercially available machine translation systems and AI agents by presenting 74 stimuli with sufficient contextual information to disambiguate the coreference of adjacent dative pronouns.

The results, analyzed with a logistic regression mixed-effect model, showed that AI-based translating agents performed better than translation systems. The study suggested that both the correspondence between word order and hierarchical prominence play roles in solving the task, and that phenomena studied in theoretical linguistics and typology are indeed retrievable from the linguistic behavior of LLMs.

Martin Meindl and **Elena Renje**, both PhD candidates at the Institute of Slavic Studies at the University of Freiburg (Germany), presented a comprehensive workflow for processing and analyzing premodern Slavic manuscripts and prints. Their approach begins with digitized texts in image format and extends to potential analysis results through several stages.

First, they addressed various aspects of automatic text recognition using the Transkribus platform, from model training to actual Handwritten Text Recognition (HTR), emphasizing the importance of training data selection. Next, they discussed post-processing generated data, including quality evaluation through Character Error Rate (CER) or Word Error Rate (WER) calculations and subsequent targeted corrections.

The workflow then prepared the expertly revised transcription for further analysis through tokenization with UDPipe and tagging with Stanzatagger. The presentation also explored the limits and possibilities of analyzing uncorrected HTR data, demonstrating that analysis of such data can yield satisfactory results with certain limitations. The discussed methods included stylometric approaches (text clustering, Nearest Shrunken Centroids algorithm) and descriptive statistical procedures.

The researchers emphasized the importance of quality control in the analysis results and advocated for a mixed-methods approach that combines quantitative robustness with qualitative assessment when working with premodern manuscripts.

Beatrice Bindi, a PhD student at the Institute of Slavic Studies at the University of Chieti-Pescara (Italy), presented research on the linguistic annotation of premodern Russian

texts, focusing on evaluating the performance of Stanza and UDPipe tools on selected works of Maximus the Greek. Her work is part of a doctoral project aiming to conduct a lexical-semantic analysis of Maximus's writings, beginning with creating a digital corpus of the author's texts from modern critical editions.

Bindi's presentation addressed the challenges of annotating historical corpora of the Russian language, noting that while specialists have developed individual solutions for linguistic annotation of ancient texts, the performance of the automatic taggers is still unsatisfying. Her qualitative evaluation of Stanza and UDPipe included the identification and analysis of common error types, providing insights into the potential future applications of these taggers for automating the linguistic annotation of Church Slavonic texts.

Fabio Maion, PhD candidate at the Institute of Slavic Studies at the University of Innsbruck (Austria), focused on the development of a historical corpus of Balkan Slavic. He highlighted that while digital methods can contribute to the study of historical language varieties, the development of diachronic corpora in Slavic studies has primarily focused on Old Church Slavonic as the oldest language stage and on texts from the East Slavic region.

Maion noted that there are currently few such resources for South Slavic language varieties. While the University of Sofia has compiled a collection of South Slavic texts from the 11th to 18th centuries, it is limited to text reproduction and only allows searches for word forms. South Slavic texts from the late Middle Ages have not yet been processed in an annotated corpus searchable by linguistic criteria, despite the rich translation literature on the Balkans in the 13th and 14th centuries and the profound structural changes that Balkan Slavic has undergone since its first attestation.

Maion discussed current projects at the University of Innsbruck's Slavic Studies department aimed at filling this gap. He presented methods for linguistically processing texts from the Second Bulgarian Empire period and outlined how an infrastructure for a historical corpus is being developed in collaboration with Bulgarian colleagues. Using the long and short forms of adjectives as a case study, he illustrated the possibilities offered by digitally annotated corpora and the new insights that can be gained from corresponding queries.

Maximilian Grübsch, PhD student at the Institute of Slavic Studies at the University of Vienna (Austria), presented research on the imperfective future tense construction *budu* + infinitive in Late Middle Russian, examining a period of profound changes in the verbal paradigm. He explained that this construction, borrowed from Polish, entered Russian during a time when a variety of infinitive periphrases referring to future time were in use, such as *stanu* and *učnu* with infinitive.

The presentation highlighted how recent scholarship has made substantial progress in understanding the semantic and syntactic distribution of these constructions, though primarily through manual annotation. Grübsch proposed applying statistical methods to verify hypotheses obtained through traditional methods and provide a broader perspective on these linguistic phenomena.

Specifically, he described using multidimensional scaling to measure distances between constructions in the usage of collostructions. Grübsch applied this approach to Late Middle Russian periphrases with future reference, modal meaning, and inchoative meaning, analyzing differences in their combinations with various infinitives.

While this method partially corroborated previous assumptions, Grübsch noted that the statistical models performed poorly, suggesting that infinitives alone don't provide sufficient context for a thorough statistical investigation of Late Middle Russian infinitive periphrases with future reference. His work illustrated both the potential and limitations of applying computational approaches to historical linguistic questions.

Ilia Afanasev, PhD student at the Institute of Slavic Studies at the University of Vienna (Austria), presented innovative research on the automatic detection of Swadesh list items from raw corpora using contemporary East Slavic standard varieties (Ukrainian, Belarusian, and Russian). He explained how this task supports the early stages of historical linguistics studies by helping researchers compile word lists for further analysis, uniquely combining semasiological and onomasiological approaches for corpus data exploration.

The central question of Afanasev's research was whether machine learning methods could detect notions of “basicness” and “swadeshness” in vocabulary. His experiments utilized the ASJP list, a well-known variation of the Swadesh list, enriched with additional words from a 110-item Swadesh list.

Afanasev detailed his methodological approach, which involved splitting the list into two halves (α-list and β-list) based on distributional semantic criteria, with words being either direct antonyms or closely semantically related. The corpus was similarly divided into three parts: α-corpus (containing only clauses with α-list items), β-corpus (with only β-list items), and a neutral part without any Swadesh list items.

For classification, Afanasev employed several methods robust enough to handle the heavily skewed classes in the dataset: Random Forest as a baseline, alongside hybridized Hidden Markov Model (h-HMM), Conditional Random Fields (CRF), and mBERT with NER head.

The results showed that while the models performed relatively poorly from a quantitative perspective, the qualitative evaluation revealed more complex insights about the relationship between tokens with similar semantics across closely-related languages. Afanasev's work contributes to a better evaluation of the term "basic vocabulary" in historical comparative research and demonstrates the challenges and potential of computational approaches to historical linguistics.

It should be noted that two researchers could not attend the conference at short notice. **Vladimir Polomac**, full professor at the Faculty of Philology and Arts at the University of Kragujevac, would have covered the important subject of developing a tagset for morphosyntactic annotation within UD for Serbian medieval charters and letters as part of creating an electronic corpus of these historical texts. **Vladimir Neumann**, Subject Specialist in Slavistics at the Eastern Europe Department of the Berlin State Library, would have presented on the effective use of NLP and search engine technologies for analyzing diachronic Slavic texts, particularly Old Church Slavonic textual corpus materials and historical Slavic dictionaries. Both these contributions, while absent from the in-person event, will be represented in the conference proceedings, which are to appear in the journal *Scripta & e-Scripta* (published by the Bulgarian Academy of Sciences).

The two-day event covered a broad spectrum of topics in both linguistic and methodological terms, addressing both synchronic and diachronic research questions, and comprising innovative methodologies and tools that are advancing our understanding of linguistic phenomena of Slavic languages across historical periods. The presentations revealed both the progress made in developing digital resources for Slavic languages and the ongoing challenges, particularly for low-resource languages and historical varieties. Furthermore, they showed that digital methods are not only being used for resource creation and workflow improvement but are increasingly being applied to test and refine theoretical linguistic hypotheses. The panel highlighted several important trends in the field, such as:

- The growing application of advanced computational methods (NLP, AI) to both contemporary and historical Slavic linguistic research
- Efforts to create much-needed digital resources for under-represented Slavic languages
- The development of specialized workflows and methodologies for processing and analyzing historical Slavic texts
- Critical evaluation of computational tools' performance on Slavic language materials

- The productive combination of quantitative and qualitative approaches through mixed-methods research
- The application of statistical modeling and machine learning techniques to address specific theoretical questions in historical and comparative Slavic linguistics

The panel concluded with a final discussion, during which participants discussed which aspects still require further development and planned to maintain the format of the DH group within Slavic studies, with such meetings to be held regularly. The organizers intend to apply for a panel at the upcoming ÖLT 2025 as well.