Proceedings of the International Conference South Slavic Languages in Digital Environment – JuDig, Vol. I/1: Thematic Collection of Papers; editors Jasmina Moskovljević Popović and Ranka Stanković. Belgrade : University, Faculty of Philology, 2025, 504 pp.

Milica Dinić Marinković

milica.dinic.marinkovic@fil.bg.ac.rs
ORCID: 0000-0002-2641-8806

University of Belgrade
Faculty of Philology
Department of General Linguistics
Belgrade, Serbia

PAPER SUBMITTED:
30 December 2025

PAPER ACCEPTED:
31 December 2025

The thematic collection of papers *Proceedings of the International Conference South Slavic Languages in Digital Environment – JuDig, Vol. I/1*[1] was published in December, 2025, within the series Scientific Conferences of the Faculty of Philology, University of Belgrade. It is currently available on the official conference website[2] and permanently archived in the DOI repos-

[1] Proceedings of the International Conference South Slavic Languages in the Digital Environment JuDig : Thematic Collection of Papers / [ed. Jasmina Moskovljević Popović, Ranka Stanković]. - Belgrade : University of Belgrade — Faculty of Philology, 2025. - ISBN 978-86-6153-791-2.

[2] Proceedings of the International Conference South Slavic Languages in the Digital Environment JuDig

itory of the Faculty of Philology, University of Belgrade[3]. The Collection contains 29 selected papers, originally presented at the international scientific conference *South Slavic Languages in Digital Environment – JuDig*, held from November 21 to 23, 2024, and co-organized by the Faculty of Philology and the Language Resources and Technologies Society– JeRTeh at the Faculty of Philology, University of Belgrade. The publication constitutes a clearly defined thematic unit regarding the contemporary research in language technologies, corpus linguistics, digital humanities, and natural language processing (NLP), with particular emphasis on South Slavic languages, while simultaneously maintaining an explicit openness toward a broader, international and multilingual research context.

The very structure of the Collection testifies to the interdisciplinary character of contemporary linguistic research in natural language processing. The selected scientific papers integrate theoretical linguistic issues with methodological and technological solutions, drawing on current approaches in computer science, artificial intelligence, and large-scale language data analysis. The thematic framework of the publication encompasses key segments of modern computational linguistics, including the construction and evaluation of language corpora, systems of linguistic annotation, the development and application of language models, speech and text processing, as well as the use of digital language resources in education, lexicography, and the preservation of linguistic and cultural heritage.

A particularly significant thematic axis of the Collection consists of papers concerning the corpora as the fundamental infrastructure of contemporary NLP research. In the paper *Quality Textual Corpora and New South Slavic Language Models* (Mihailo Škorić, Saša Petalinkar), the relationship between the quality of textual corpora and the performance of modern language models is examined, with special attention to the challenges posed by South Slavic languages in the context of large language models. The paper *Syntactic-Semantic Annotation of the Electronic Corpus of the Serbian Language* (Nataša Kiš) addresses issues of designing and implementing complex annotation systems, highlighting the importance of precise and theoretically grounded annotation for further automatic processing and linguistic analysis.

The topic of parallel corpora and their research potential for contrastive studies, machine translation improvement, translation theory, and the investigation of specialized genres is also well represented. In the paper *The Importance of Parallel Corpora for Research on Phraseological Constructions*

[3] doiFil/Faculty of Philology, University of Belgrade. Proceedings of the International Conference South Slavic Languages in the Digital Environment JuDig

in German and Serbian (Kristina Ilić), parallel corpora are viewed as a key resource for contrastive research, particularly in the domain of phraseology and lexical semantics. A similar methodological framework, but with a focus on the development of specialized multilingual resources, is presented in the paper *Methodology for Building a Multilingual Parallel Corpus Based on Online Digital User Manuals: The Hilti Manuals Corpus* (Nikola Janković), which details the stages of data collection, processing, and alignment in contemporary parallel corpora.

Another major thematic unit comprises papers oriented toward the development and evaluation of tools and methods for automatic language processing. In the paper *Exploring the Synergy Between LLMs and Knowledge Graphs for Advanced Abusive Speech Detection in Serbian* (Danka Jokić, Ranka Stanković), the integration of large language models and knowledge graphs is examined in tasks of abusive speech detection, linking modern machine-learning approaches with symbolic knowledge models. The paper *New Improved Versions of the Program "Towards Minimal Pairs"* (Danilo Aleksić) presents enhancements to existing tools for phonological and phonetic analysis, demonstrating how classical linguistic issues can be effectively connected with modern software solutions.

Speech and multimodal data processing are explored in the paper *The Use of the Whisper Large v3 Sr Model for Serbian Speech Transcription in Python on the Google Colab Platform* (Nikola Janković, Jovana Ivaniš), which focuses on the practical application of contemporary ASR models in research and educational contexts. In a related but typologically broader framework, the paper *Noun Phrase and Prepositional Phrase Chunking for the Greek Language with spaCy* (Nikitas N. Karanikolas) illustrates how modern NLP tools can be adapted to languages with different morphosyntactic characteristics, further emphasizing the international dimension of the Collection.

Particular attention is also devoted to contributions that connect language technologies with lexicography and cultural heritage. In the paper *The Srpko Corpus as a Technological Basis for the Dictionary of Contemporary Serbian language of Matica srpska* (Dušanka Vujović, Branko Milosavljević), the corpus is considered a foundation of modern lexicographic practice, while the paper *From the Dictionary as Corpus to Hyponymy and Meronymy* (Maja Matijević) explores the relationships between lexicographic resources and semantic relations, pointing to the potential of the corpus as a source for both theoretical and applied lexical semantics.

The broader social and educational context of application of language technologies is discussed in the paper *Integration of the Serbian Language Version of Wikipedia into Educational Systems and the Advancement of Language Technologies* (Nebojša Ratković), which analyzes the role of open digital resources in education and in the development of language technologies. Such works link the technical aspects of language processing with issues of knowledge accessibility, standardization, and institutional support in the development of language resources.

The Collection may also be viewed as a document of a particular historical moment in which South Slavic languages are consolidating their position and increasing their visibility on the map of contemporary research in language technologies. The papers in this Collection not only present existing achievements, but also clearly indicate directions for future development, the need for sustainable research infrastructure, and the importance of continuous inter-institutional and international cooperation. In this sense, *Proceedings of the International Conference South Slavic Languages in Digital Environment – JuDig, Vol. I/1* positions itself as a publication that both documents and actively builds a contemporary research network in the fields of language technologies and computational linguistics across the South Slavic area. The special value of the Collection is reflected in the visible and substantively grounded cooperation of researchers from Serbia, Croatia and Slovenia, who, through joint projects, tools and resources contribute to the development of an interoperable and sustainable infrastructure for the research of South Slavic languages in a digital environment.

This collaborative aspect is particularly evident in papers dedicated to the development of shared resources and methodologies for South Slavic languages. In addition to the paper *Quality Textual Corpora and New South Slavic Language Models* (Mihailo Škorić, Saša Petalinkar), which addresses issues of corpus construction and evaluation as a shared foundation for language modeling of South Slavic languages, a significant contribution to regional collaboration is also provided by the paper *Models for Automatic Morphological Inflection of Serbian and Croatian Based on the srLex and hrLex Morphological Lexicons* (Jaka Čibej). In this work, starting from the existing morphological lexicons, *srLex* and *hrLex*, models for automatic inflectional morphology of Serbian and Croatian are developed, drawing on the experience and tools originally developed for Slovenian language. This clearly demonstrates how knowledge and resources developed for one language can be systematically transferred and adapted to related languages, while ensuring open access and availability of the results.

Another prominent example of such collaboration is the paper *The ELEXIS-WSD Parallel Sense-Annotated Corpus and South Slavic Languages: Subcorpora for Croatian, Serbian, and Slovene* (Jaka Čibej, Ranka Stanković, Ana Ostroški Anić, Simon Krek, Carole Tiberius), which describes the expansion of the parallel sense-annotated ELEXIS-WSD corpus within the framework of the ELEXIS project and the COST Action UniDive. The focus on the Croatian, Serbian, and Slovenian subcorpora clearly illustrates the efforts to create high-quality, manually annotated resources that are both methodologically harmonized and applicable in various NLP tasks, including word-sense disambiguation. Such resources have multiple significances: they connect the lexicographic traditions of related languages, enable contrastive and multilingual research, and provide a foundation for further technological development.

The regional dimension of the collaboration is further reinforced by the papers linking South Slavic languages in the domains of phonology and language processing, such as the paper *Sonority Based Syllabic Hyphenation of Macedonian and Serbian* (Katerina Zdravkova, Jana Kuzmanova), where a shared methodological framework is applied to genealogically related languages, highlighting the potential for broader regional integration.

On a distinct thematic level, the paper *On the Family of Contemporary Serbian Language Corpora SrpKor* (Duško Vitas, Ranka Stanković, Cvetana Krstev) provides a valuable overview of the development of computational linguistics in Serbia through the lens of the evolution of one of its most important language resources. The presentation of SrpKor's development—from changes in tools and corpus dimensions, through the expansion of temporal and genre coverage, to increasingly complex levels of annotation and linking with lexical resources via the Leksimirka system—offers not only a historical overview, but also a clear insight into how national research infrastructure and computational support have developed along with international scientific trends.

The exceptional value of the publication lies in the balance it achieves between highly topical research themes in the field of natural language processing, such as large language models, machine learning, and automatic speech processing, and studies aimed at the long-term development of language resources, lexicography, and linguistic heritage. Viewed as a whole, the Collection successfully integrates current "cutting-edge" technological approaches with research focused on preservation, documentation, and systematization of linguistic data, offering theoretical, methodological, and practical insights simultaneously. A further significant merit of the publication is its role in connecting researchers from diverse disciplines and institutions, actively building a network of experts dedicated to the development of language resources and technologies for South Slavic languages. Therefore, *Proceedings of the International Conference South Slavic Languages in Digital Environment – JuDig, Vol. I/1* transcends the boundaries of a conventional conference collection and establishes itself as a relevant reference point for understanding both the current state and future directions of language technologies and computational linguistics in the South Slavic and broader European context.