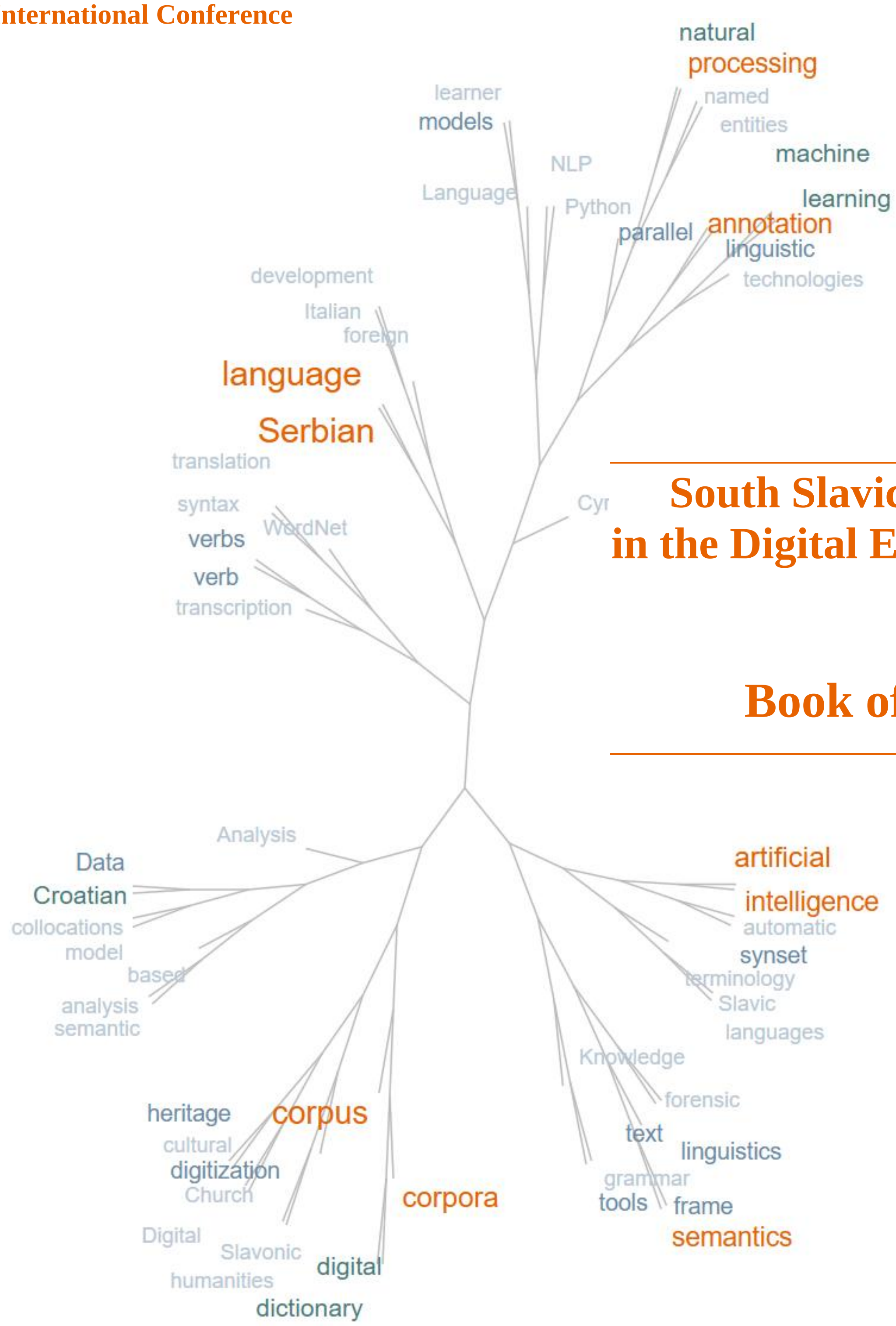
International Conference



South Slavic Languages in the Digital Environment

JuDig

Book of Abstracts

November 21-23, 2024.
University of Belgrade - Faculty of Philology, Serbia

ORGANISERS
University of Belgrade - Faculty of Philology
Society for Language Resources and Technologies JeRTeh

Editors:
Prof. dr Jasmina Moskovljević Popović
Prof. dr Ranka Stanković

Technical editors:
Prof. dr Ranka Stanković
Dr Aleksandra Tomašević

Graphic design:
Anđelka Zečević

Publisher:
University of Belgrade - Faculty of Philology

Publication accessible at:
https://judig.jerteh.rs/2024/judig-book-of-abstracts.pdf

Printed by
University of Belgrade - Faculty of Philology

Print run:
100

Website: https://judig.jerteh.rs/
E-mail: judig@jerteh.rs
Graphics: https://treecloud.univ-mlv.fr/

ISBN-978-86-6153-754-7

The conference is supported by

INSTITUTE FOR
STANDARDIZATION
OF SERBIA

(bronze sponsor)

Science Fund
of the Republic of Serbia

Republic of Serbia
MINISTRY OF SCIENCE,
TECHNOLOGICAL DEVELOPMENT AND INNOVATION

Office for Information
Technologies and e-Government

University of Belgrade
Faculty of Mining and Geology

ENVI SOFTWARE SOLUTIONS

PROGRAMME COMMITTEE

Co-Chairs:

Prof. dr Jasmina Moskovljević Popović, University of Belgrade, Faculty of Philology
Prof. dr Ranka Stanković, University of Belgrade, Faculty of Mining and Geology and Society for Language Resources and Technologies JeRTeh

Members:

Prof. dr Agata Savary, University of Paris-Saclay, France
Dr Aleksandra Marković, Institute for Serbian, SANU, Serbia
Dr Ana Ostroški Anić, Institute of Croatian Language, Croatia
Doc. dr Balša Stipčević, University of Belgrade - Faculty of Philology, Serbia
Prof. dr Benedikt Perak, University of Rijeka, Faculty of Philosophy, Croatia
Dr Biljana Rujević, University of Belgrade, Faculty of Mining and Geology, Serbia
Dr Vasilije Milnović, University Library "Svetozar Marković", Serbia
Prof. dr Vera Ćevriz Nišić, University of East Sarajevo, Faculty of Philosophy, Bosnia and Herzegovina
Dr Verginica Barbu Mititelu, Research Institute for Artificial Intelligence, NLP Group, Romanian Academy, Romania
Prof. dr Vladan Devedžić, University of Belgrade, Faculty of Organizational Sciences, Serbia
Prof. dr Vladimir Polomac, University of Kragujevac, Faculty of Philology and Arts, Serbia
Prof. dr Gordana Pavlović-Lažetić, Language Resources and Technologies Society (JeRTeh), Serbia
Doc. dr Danilo Aleksić, University of Belgrade, Faculty of Philology, Serbia
Prof. dr Dimitar Trajanov, Saints Cyril and Methodius University - Skopje, Faculty of Computer Science and Engineering, Macedonia
Prof. dr Dušanka Popović, University of Montenegro, Faculty of Philosophy, Montenegro
Prof. dr Duška Klikovac, University of Belgrade - Faculty of Philology, Serbia
Prof. dr Duško Vitas, Language Resources and Technologies Society (JeRTeh), Serbia
Dr Jaka Čibej, University of Ljubljana, Faculty of Philosophy, Slovenia
Prof. dr Jelena Graovac, University of Belgrade, Faculty of Mathematics, Serbia
Prof. dr Jelena Jovanović, University of Belgrade, Faculty of Organizational Sciences, Serbia
Prof. dr Jelena Marković, University of East Sarajevo, Faculty of Philosophy, Bosnia and Herzegovina
Doc. dr Jovan Čudomirović, University of Belgrade - Faculty of Philology, Serbia
Prof. dr Katerina Zdravkova, Saints Cyril and Methodius University - Skopje, Faculty of Computer Science and Engineering, Macedonia
Prof. dr Marko Robnik Šikonja, University of Ljubljana, Faculty of Computer and Information Science, Slovenia
Doc. dr Miloš Utvić, University of Belgrade - Faculty of Philology, Serbia
Dr Mihailo Škorić, University of Belgrade, Faculty of Mining and Geology, Serbia
Dr Nebojša Vasiljević, Petlja Foundation, Serbia
Prof. dr Nevena Ceković, University of Belgrade - Faculty of Philology, Serbia
Prof. dr Nelda Kote, Polytechnic University of Tirana, Albania
Prof. dr Olja Perišić, University of Turin, Department of Foreign Languages, Literatures and Modern Cultures, Italy

Dr Paraskevi Giouli, Institute for Language and Speech Processing, Athens Research Center, Greece
Prof. dr Petja Osenova, Sofia University "St. Kliment Ohridski", Bulgaria
Prof. dr Rada Stijović, Language Resources and Technologies Society (JeRTeh), Serbia
Doc. dr Saša Marjanović, University of Belgrade - Faculty of Philology, Serbia
Prof. dr Saša Moderc, University of Belgrade - Faculty of Philology, Serbia
Prof. dr Svetla Koeva, Bulgarian Academy of Sciences, Bulgaria
Prof. dr Sonja Nenezić, University of Montenegro, Faculty of Philologyž, Montenegro
Prof. dr Cvetana Krstev, Language Resources and Technologies Society (JeRTeh), Serbia
Dr Kristina Štrkalj Despot, Institute of Croatian Language, Croatia
Prof. dr Ivan Obradović, Language Resources and Technologies Society (JeRTeh), Serbia
Dr Olivera Kitanović, University of Belgrade, Faculty of Mining and Geology, Serbia

ORGANISING COMMITTEE

Co-Chairs:

Prof. dr Jasmina Moskovljević Popović, University of Belgrade, Faculty of Philology
Prof. dr Ranka Stanković, University of Belgrade, Faculty of Mining and Geology and Society for Language Resources and Technologies JeRTeh

Members:

Doc. dr Jovan Čudomirović, Faculty of Philology, University of Belgrade
Doc. dr Balša Stipčević, Faculty of Philology, University of Belgrade
Prof. dr Nevena Ceković, Faculty of Philology, University of Belgrade
Doc. dr Miloš Utvić, Faculty of Philology, University of Belgrade
Doc. dr Milica Dinić Marinković, Faculty of Philology, University of Belgrade
MSc Milica Ikonić Nešić, Faculty of Philology, University of Belgrade
Dr Biljana Rujević, Language Resources and Technologies Society (JeRTeh)
Dr Mihailo Škorić, Language Resources and Technologies Society (JeRTeh)
Dr Aleksandra Marković, Language Resources and Technologies Society (JeRTeh)
Anđelka Zečević, Language Resources and Technologies Society (JeRTeh)
Nikola Gudžić, Language Resources and Technologies Society (JeRTeh)

The International Conference South Slavic Languages in Digital Environment—JuDig provides an opportunity for numerous researchers in the field of computational linguistics, language technologies, and related fields to share their ideas, insights, and results in these research areas.

The aim of the conference is to foster networking and to provide a forum for high-quality research in all (sub)domains of computational linguistics and related topics, with a special focus on language resources and technologies available for Serbian and other South Slavic and Balkan languages. Forty-three eminent scientists, coming from twelve countries and twenty-two research institutions participate as members of the Conference Programme Committee.

Fifty-five contributions on a large number of topics in the field of language technologies were submitted for the conference. All submissions underwent a double-blind peer review process by three members of the Programme Committee. The Programme Committee and the co-chairs selected 51 papers written by 84 (co)authors from 34 institutions and 14 countries to be presented at the conference – fifty-five from Serbia, five from Croatia, four from Bulgaria and Germany, two from Slovenia, Austria, Macedonia, United Kingdom, France, and Bosnia and Herzegovina, and one from Greece, the Netherlands, South Korea, and Italy. The Conference is held in a hybrid mode and includes four invited talks, twelve parallel sessions and four thematically oriented workshops.

The esteemed invited speakers will present on pertinent topics and recent innovations in language technologies, offering insights into cutting-edge advancements and significant trends which shape the discipline. They come from four countries: France (Agata Savary), Bulgaria (Svetla Koeva), United Kingdom (Ruslan Mitkov) and Serbia (Cvetana Krstev).

The field of language technologies is nowadays not only very popular, but also very broad. As a result, a wide range of topics will be tackled at the conference. The 51 accepted abstracts have been classified into seven topics: IT processing of the South Slavic languages (5), Digital corpora of the South Slavic languages (11), Language resources for the South Slavic languages (6), Language technologies for the South Slavic languages (8), Grammar and lexicon of the South Slavic languages in the context of NLP (5), Artificial intelligence, language models, and the processing of the South Slavic languages (8), and Digital humanities (8).

The Organising Committee has particularly focused on sharing and promoting the existing resources and technologies with other scientists and prospective researchers. In four workshops the participants will have an opportunity to learn about the latest approaches, resources, and tools in several different areas, to explore examples of good practice (short case studies) conducted by renowned experts, and to establish contacts with future collaborators.

The organisers extend their sincere gratitude to the sponsors and supporters of the Conference. Their generous contributions and commitment have been instrumental in making this event possible. We believe that with joint efforts we could make a considerable impact on the future of research in this crucial scientific area.

Table of Contents

Invited Talks

Contributed talks

IT Processing of South Slavic Languages

Digital Corpora of South Slavic Languages

Language Resources for South Slavic Languages

Language Technologies for South Slavic Languages

Grammar and Lexicon of South Slavic Languages in the Context of NLP

Artificial Intelligence, Language Models And Processing Of South Slavic Languages

Digital Humanities

JUDIG Workshops

Invited Talks

Agata Savary

Professor in computer science, Université Paris-Saclay
LISN (Laboratoire Interdisciplinaire des Sciences du Numérique)
IUT d'Orsay (University Institute of Technology)
https://perso.lisn.upsaclay.fr/savary/

Agata Savary is a professor in computer science at the Paris-Saclay University in France. She holds a MSc from the University of Warsaw in Poland, a PhD from the University of Marne-la-Vallée in France, and a Habilitation from the University of Tours in France. She has been dealing with multilingual NLP for 3 decades. Her domains of interest include natural language processing (NLP), universalist and NLP-applicable modeling of idiomaticity in language, as well as construction of language resources and tools for multiword expression identification, named entity recognition, and coreference resolution.

She chairs the CA21167 COST action UniDive (Universality, diversity and idiosyncrasy in language technology, 2022-2026), which counts over 300 members from 37 countries.

In the past she also served as the chair of the IC1207 COST action PARSEME (PARSing and Multiword Expressions, 2013-2017), an elected representative of the Multiword Expressions section at SIGLEX, a co-editor-in-chief of the "Phraseology and Multiword Expressions" book series at Language Science Press, and a coordinator of the Dagstuhl Seminar on the "Universals of Linguistic Idiosyncrasy in Multilingual Computational Linguistics". She authored 20 refereed journal papers, 40 conference papers, 20 workshop papers, and 10 book chapters.

Automatic Identification of Multiword Expressions – Recent Progress and Perspectives

Multiword expressions, like "all of a sudden", a "hot dog", or to "pull one's leg", are combinations of words which exhibit idiosyncratic behavior on the lexical, morphological, syntactic, semantic, pragmatic and/or statistical levels. Their semantic non-compositionality is their most outstanding feature and can pose severe challenges in semantically-oriented NLP tasks.

One of the ways to tackle this challenge is to identify MWEs in running text before applying dedicated treatment to them.

MWE identification has been the object of many efforts, notably within the PARSEME shared task on automatic identification of verbal MWEs. I will summarize major findings from these shared tasks and highlight the particular properties of MWE which make their identification particularly challenging, even with deep learning methods.

I will also touch upon the latest challenges and opportunities from MWE processing applied to more generic NLP tasks such as neural machine translation or interpretation of neural models.

Svetla Koeva

Professor at the Bulgarian Academy of Sciences
Director of the Institute for Bulgarian Language
Chair of the Department of Computational Linguistics
http://dcl.bas.bg/news/svetla-koeva/

Dr Svetla Koeva is a professor of Computational Linguistics and the head of the Department of Computational Linguistics at the Institute for Bulgarian Language, Bulgarian Academy of Sciences. Her research interests lie in the field of computational linguistics, formal description of language: morphology and syntax, lexical-semantic networks and ontologies. She is the principal investigator in the development of various language resources for Bulgarian such as: Bulgarian WordNet, Bulgarian National Corpus, text processing chain, etc.

Svetla Koeva has published 5 books and over 200 research publications. She has lead numerous research projects among which currently are: Assessing reading literacy and comprehension of early graders in Bulgaria and Italy and Enriching the Semantic Network WordNet with Conceptual Frames, and the successfully finished projects in the last two years are: European Lexicographic Infrastructure, Multilingual Image Corpus (MIC 21), The ontology of the stative situations in the models of language, Curated Multilingual Language Resources for CEF.AT. Since 2013 Svetla Koeva leads the project The Written Word Remains. Write Correctly! for promoting the study and research of the Bulgarian language. She also is the head of the National Competence Centre at the European Language Grid.

She served as the Director of the Institute for Bulgarian Language from 2012 until 2021. Currently she is the chair of the Research Council of the Institute for Bulgarian Language since 2021. Svetla Koeva is also the editor in chief of the Annual Papers of the Institute for Bulgarian Language and the weekly edition The Written Word Remains. Write Correctly!. Svetla Koeva has been awarded five prizes by the National Science Fund at the Ministry of Education and Science and the Bulgarian Academy of Sciences.

LLM-Based (Computational) Linguistics Research: Between Profit and Risk

The field of computational linguistics has undergone major conceptual changes, moving from symbolic techniques to machine learning and deep learning. Large Language Models (LLMs) are now replacing many traditional NLP technologies in various application areas such as question answering, summarization, text simplification, and named entity recognition. Moreover, LLMs are proving to be capable of data analysis within a given theoretical framework, frame semantics research, and corpus-based linguistic studies by automatically annotating texts with specific linguistic information.

This talk aims to discuss the challenges of using LLMs in (computational) linguistics research, drawing on existing and ongoing studies. We will explore the benefits, challenges, limitations, and potential risks associated with the use of LLMs for research, with a particular emphasis on the South Slavic languages in comparison to English. Through this analysis, we aim to provide insights into the application of LLMs in computational linguistics and identify areas for further investigation and improvement.

Ruslan Mitkov

Professor of Computing and Communications, School of Computing and Communications, Lancaster University
Professor of Computational Linguistics and Language Engineering
Director of the Erasmus Mundus Programme on Technology for Translation and Interpreting (EM TTI)
Executive Editor of the journal 'Natural Language Processing' (Cambridge University Press)
Programme Chair of the RANLP conference series
https://wp.lancs.ac.uk/mitkov

Prof Dr Ruslan Mitkov is Professor in Computing and Communications at Lancaster University, one of the top-10 UK universities. Prior to joining Lancaster University, Prof Mitkov worked at the University of Wolverhampton where he created and led the internationally leading Research Group in Computational Linguistics, and was also Director of the Research Institute of Information and Language Processing as well as Director of the Responsible Digital Humanities Lab. Prof Mitkov is also Distinguished Professor at the University of Alicante, Spain.

Dr Mitkov has been working in Natural Language Processing (NLP), Computational Linguistics, Corpus Linguistics, Machine Translation, Translation Technology and related areas since the early 1980s. Whereas Prof Mitkov is best known for his seminal contributions to the areas of anaphora resolution, automatic generation of multiple-choice tests and new generation translation memory systems, his extensively cited research (more than 320 publications including 15 books, 35 journal articles and 35 book chapters) also includes but is not limited to topics such as computational phraseology, machine translation, natural language processing for language disabilities, automatic summarisation, computer-aided language processing, corpus annotation, bilingual term extraction, automatic identification of cognates and false friends, NLP-driven corpus-based study of translation universals and text simplification. His recent research includes the employment of Deep Learning, Large Language Models and Artificial Intelligence in Natural Language Processing, Translation Technology, Linguistics and language research in general. Prof Mitkov is not only known for his original research outputs with high scientific impact, but also known for his vision and innovative applied research which seeks to enhance the work efficiency of different professions (teachers, translators and interpreters) or seeks to improve the quality of life (people with disabilities).

Prof Mitkov is author of the monograph Anaphora resolution (Longman) and sole Editor of The Oxford Handbook of Computational Linguistics (Oxford University Press) which has been hailed as the most successful Oxford Handbook and whose second, substantially revised edition was published in June 2022. Current prestigious projects include his role as Executive Editor of the journal Natural Language Processing (formerly Journal of Natural Language Engineering) published by Cambridge University Press, Editor-in-Chief of the Natural Language Processing book series of John Benjamins publishers, and Consulting Editor of Oxford University Press publications in Computational Linguistics.

Prof Mitkov has been invited as a keynote speaker at more than 240 international conferences (28 keynote speeches in 2024 only) and is/has been Chair of more than 70 conferences on Natural Language Processing, Translation Technology and Applied Linguistics topics. He is editor of more than 15 volumes published by Springer and John Benjamins.

Mitkov designed and is Director of the first and only Erasmus Mundus Master’s programme in Technology for Translation and Interpreting - an innovative and inspirational programme, with a strong research focus but an equally strong emphasis on business; leading companies in the global translation and language industry participate as associated partners.

Prof Mitkov has been an external examiner of many doctoral theses and curricula in the UK and abroad, including Master’s programmes related to Natural Language Processing, Computational Linguistics, Digital Humanities, Translation and Translation Technology.

Ruslan Mitkov received his MSc from the Humboldt University in Berlin, his PhD from the Technical University in Dresden and worked as a Research Professor at the Institute of Mathematics, Bulgarian Academy of Sciences, Sofia. He is a Fellow of the Alexander von Humboldt Foundation, Germany, Marie Curie Fellow, Distinguished Visiting Professor at the University of Franche-Comté in Besançon, France and Distinguished Visiting Researcher at the University of Malaga, Spain.

Ruslan Mitkov is Vice President of AsLing an international Association for promoting Language Technology. In September 2022 the renowned National Board of Medical Examiners (USA) presented Prof Mitkov with a certificate of distinguished collaboration which resulted in lasting impact on the strategic planning and decision making of the US organisation and their employment of NLP solutions to assessment for the last 17 years. In recognition of his outstanding professional/research achievements, Prof Mitkov was awarded the title of Doctor Honoris Causa three times.

The Evolution of Natural Language Processing: from Rules Through Neural Networks to Generative AI. What Does the Future Hold?

Natural Language Processing (NLP) is undergoing dynamic and unprecedented changes as never before. While we have always known that NLP is not a magic technology which always has been far from 100% accurate, the landscape of Language and Translation Technology is changing. First Deep Learning methods and now Large Language Models, have taken the world by storm. This easy-to-follow and entertaining talk will seek to shed light on the future of Natural Language Processing in the Artificial Intelligence (AI) era.

The keynote will sketch the history of Natural Language Processing and Machine Translation and will review the latest advances powered by Deep Learning and Large Language Models (LLMs). It will then critically look at the employment of LLMs in Natural Language Processing and Machine Translation (and the evolution of NLP methods will be exemplified by) reporting on recent original research of the speaker which compares LLMs, Deep Learning, and rule-based approaches for selected NLP tasks and applications.

These studies will serve as a platform for a follow-up discussion on the future of Natural Language Processing. The speaker will emphasize that he is not a clairvoyant but based on his experience in the field, he will attempt to predict the likely future of artificial intelligence as compared to human intelligence taking language as a testbed.

Cvetana Krstev

Professor of Informatics at the University of Belgrade - Faculty of Philology (retired)
President of the Language Resources and Technologies Society (JeRTeh)
https://poincare.matf.bg.ac.rs/~cvetana/

Prof. Dr Cvetana Krstev is a retired professor of Informatics at the Department of Librarianship and Informatics, University of Belgrade - Faculty of Philology. Her scientific field is Human Language Technologies (HLT). She has published one book and more than 200 scientific papers, most related to natural language processing, more specifically to language resources development and their application. She has developed the Serbian morphological e-dictionary and was one of the key contributors to the development of the Serbian WordNet, Serbian monolingual and aligned multilingual corpora, the named-entity recognition system for Serbian, as well as systems for automatic text correction, for various text transformations, corpus-lexicon interaction and many other language resources and tools.

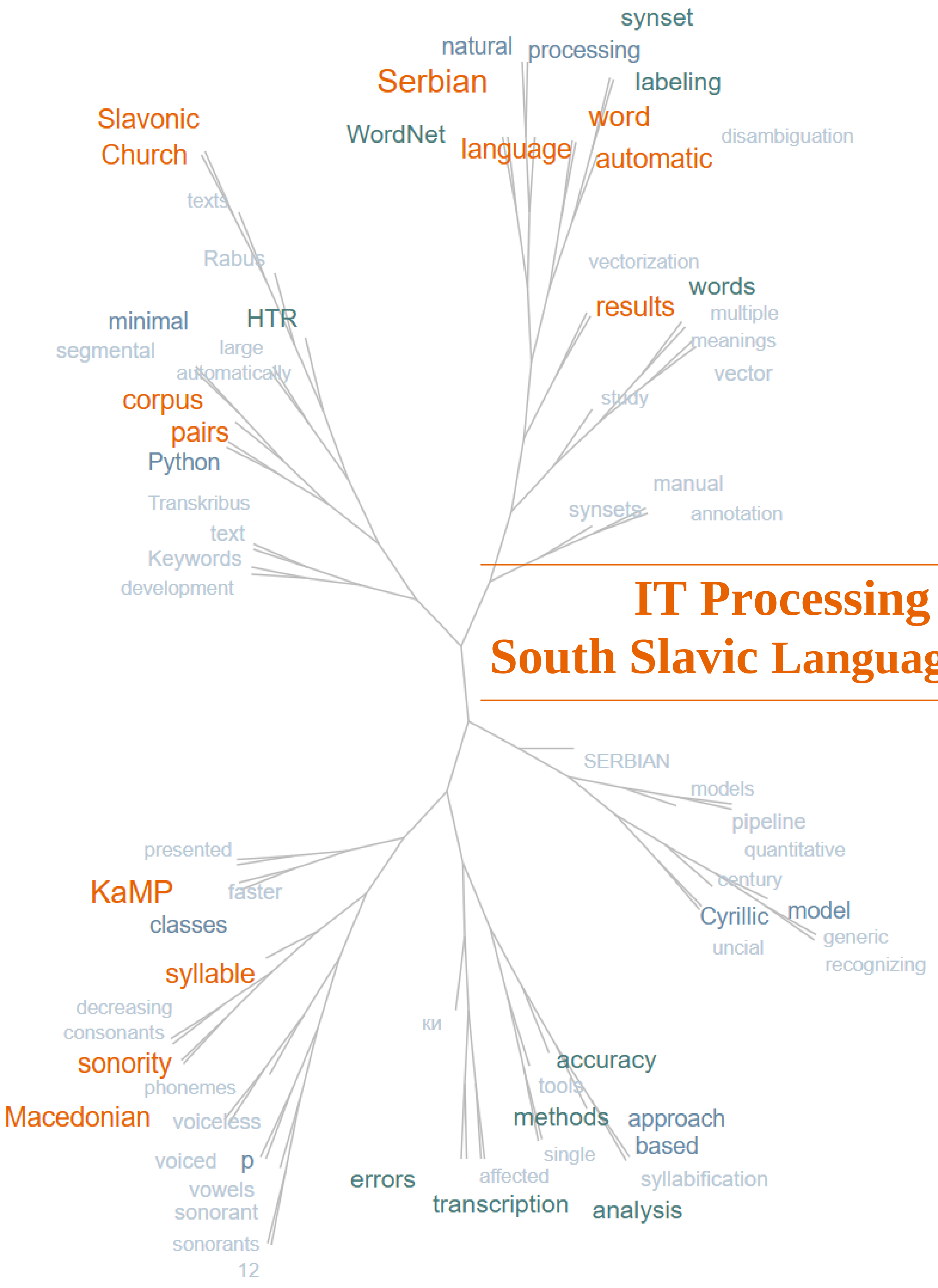
She participated in several international and national projects related to language resources and technology. Also, she is the president and one of the initiators and founder-members of the Language Resources and Technologies Society (JeRTeh). She also is the head of the National Competence Centre at the European Language Grid. She was MC and a significant contributor to the COST actions: PARSEME (PARSing and Multiword Expressions, 2013-2017), D-Reading (Distant Reading for European Literary History, 2017-2021) and currently leading activities as MC in UNIDIVE COST action (Universality, diversity and idiosyncrasy in language technology, 2022-2026).

Relation Between „Us“ and „Others“ in the SrpELTeC Corpus Retrieved Using Good Old-Fashioned Methods

SrpELTeC corpus is a part of a large multilingual collection of novels written in the period 1840-1920. It contains 120 novels written originally in the Serbian language. Due to its precise annotations, it was already used for research in various domains: linguistic, philological, and cultural. In the present research, we will try to establish how the relation between „us“ (Serbians by nation or citizenship) and „others“ is reflected in novels written in the second half of the 19th century and the beginning of the 20th century. We will use annotations introduced manually, semi-automatically, or automatically, the last two relying on comprehensive lexical resources and tools based on them. We will try to show that the good old methods, such as reading and counting, can still produce some interesting results.

Contributed talks

IT Processing of South Slavic Languages



Saša Petalinkar
University of Belgrade, JeRTeh
E-mail: sasa5linkar@gmail.com

Automatic Synset Labeling for Serbian Wordnet: Development, Preprocessing, and Feasibility Assessment

This research presents the development of an annotator for the Serbian WordNet, designed to assign appropriate synsets to each word in a given text. The core framework is based on the spaCy pipeline enhanced with the BERT Jerteh 355 model, enabling various natural language processing tasks, including lemmatization. A custom layer is integrated to facilitate synset labeling.

Due to the complexity of lemmatizing n-grams, this study focuses on labeling unigrams with their corresponding synsets. The main challenge addressed is word sense disambiguation for words with multiple meanings. The goal of this study is to assess the feasibility of an automatic solution using embeddings and vector similarity between words and synset descriptions, or the necessity for manual annotation to build a training corpus.

The benefit of the WordNet annotator is that it enables accurate and rapid labeling of word meanings, which is essential for natural language processing and numerous applications in artificial intelligence. WordNet provides a structured database that links words with their meanings and relations, enhancing the understanding of context and semantics.

As a preprocessing step, all definitions from the Serbian WordNet are vectorized. In the relevant layer, the lemma obtained from the previous layers of the spaCy pipeline is used to search for all synsets containing that word. If there is one or no synset, it is automatically added; if there are multiple, a vector comparison is performed between the word's vector and the already vectorized definitions. The study explores which types of vectorization are possible and what results they yield.

Preliminary results have shown that vectorization using SrpCNNER yields the following results when applied to twenty sentences from a parallel English-Serbian corpus manually annotated with WordNet: 58 correct and 87 incorrect.

In addition to automatic labeling, tools are being developed to assist annotators by suggesting possible synsets, allowing for manual annotation and the construction of a training corpus. These tools enable the evaluation of how well vector comparison can achieve good results or if further improvement through manual annotation is necessary.

These findings will contribute to the further development of tools for the Serbian language, providing insights into the potential and limitations of automatic synset labeling and word sense disambiguation. Significant implications are expected for enhancing the understanding and processing of natural language in Serbian.

Keywords: *automatic synset labelling, Serbian WordNet, word vectorization, natural language processing, synset disambiguation, BERT model*

Vladimir Polomac, Tamara Lutovac Kaznovac, Marko Milošević, Ana Marija Pavlović
Univerzitet u Kragujevcu, FILUM, Katedra za srpski jezik
E-mail: {v.polomac|tamara.kaznovac| marko.milosevic| ana.marija.pavlovic|@filum.kg.ac.rs

Models for Automatic Recognition of Old Serbian Printed and Manuscript Cyrillic: Current State and Future Tasks

This paper provides an overview of models for automatic recognition of old Serbian Cyrillic developed within the Transkribus software platform as part of the international bilateral project "Creating AI Models for the Automatic Processing of Serbian Medieval Manuscripts" between the Department of Serbian Language at the University of Kragujevac and the Institute of Slavic Studies at the University of Freiburg. The review covers: 1) a large generic model for recognizing Serbian Church Slavonic printed and manuscript Cyrillic (uncial, semi-uncial and cursive) from the 12th to the 18th century; 2) a generic model for recognizing Serbian Church Slavonic and Serbian semi-uncial and cursive Cyrillic in the manuscripts of Gavrilo Stefanović Venclović (18th century); and 3) a generic model for recognizing Serbian diplomatic Cyrillic from the 13th to the 16th century. The paper presents the quantitative and qualitative performances of these models, as well as potential directions for their further development.

Keywords: *automatic text recognition, Transkribus, artificial intelligence, machine learning, old Serbian Cyrillic*

Dr. Anna Jouravel, Martin Meindl, Prof. Dr. Achim Rabus, Elena Renje
University of Freiburg
E-mail: {anna.jouravel, martin.meindl, achim.rabus, elena.renje}@slavistik.uni-freiburg.de

Creating a Multi-Layered Digital Framework for Analysing Church Slavonic Manuscripts

In this contribution, we will discuss various digital methods for analysing Old Church Slavonic as well as Church Slavonic texts in their East and South Slavonic recensions. For the analysis, we employ a multi-layered approach using various tools and methods rather than focusing on a single approach. We believe this strategy will yield more valid and robust results, as it is less likely that all the methods combined will be affected by the errors present in the data compared to using one single tool.

These methods are part of the quantitative pipeline that begins with the automatic transcription of (Old) Church Slavonic manuscripts and continues with the analysis of these transcriptions, without the need for manual correction of the errors produced by the Handwritten Text Recognition (HTR) engine. This pipeline comprises multiple pre- and postprocessing steps, including automatic transcription, tokenization, parsing, tagging and statistical analysis. It involves the use of HTR engines such as the *Transkribus* platform, browser and application-based software like *UDPipe*, *Stanza*, *Voyant Tools* and *AntConc*, as well as statistical methods implemented in programming languages like *R* and *Python*.

Rabus (2019) demonstrated that HTR engines can automatically transcribe large amounts of Church Slavonic texts with a Character Error Rate (CER) of 4% or less, a result that aligns with the performance of the models we used in our experiments. However, the most frequent lexical units are far less affected by the errors, so we propose that the transcription results serve as a solid foundation for certain types of quantitative analysis and that the amount of transcribed text outweighs the relatively small number of errors. Nevertheless, using a small fragment of text we conduct a parallel qualitative analysis of the most frequent errors introduced by the HTR engine. In accordance with previous studies (Rabus 2019, Burlacu & Rabus 2021) we find that superscript characters and word separation are the two most common sources for errors in the transcription. This preemptive analysis allows us to better gauge the effect of erroneous transcription on our results. The mixed-methods approach not only ensures that the results derived from HTR data are logical and meaningful, but it has also been proven successful in the analysis of Church Slavonic texts (Rabus & Petrov 2023).

Keywords: *Mixed-Methods, Church Slavonic, Statistics, HTR*

References

[1] Burlacu, C. & Rabus, A. (2021): Digitising (Romanian) Cyrillic using Transkribus: new perspectives. Diacronica, 14, art. A196. P. 1-9.

[2] Rabus, A. (2019): Recognizing Handwritten Text in Slavic Manuscripts: a Neural-Network Approach Using Transkribus. Scripta & e-Scripta, 19. P. 9-32.

[3] Rabus, A. & Petrov, I. N. (2023): Linguistic Analysis of Church Slavonic Documents: A Mixed-Methods Approach. Scando-Slavica, 69.1. P. 25-38.

Katerina Zdravkova, Jana Kuzmanova
Faculty of Computer Science and Engineering, Skopje
E-mail: {katerina.zdravkova; jana.kuzmanova}@finki.ukim.mk

Sonority Based Syllabification of Macedonian and Serbian Language

Phonetically, syllables are sequences of sounds that contain a single peak of prominence, while phonologically they are units of stress placement. According to the Sound Sequencing Principle, sonority within a syllable rises to the nucleus of the syllable and then falls in sonority. So far, there were several attempts to syllabify Macedonian and Serbian words. The accuracy of the Macedonian experiment was not evaluated on a specific corpus, while the Serbian syllabification exceeded 98%. The rule-based approach was rather complex, compared to sonority based syllabification that we proposed for Macedonian and extend to Serbian.

The sonority of Macedonian phonemes depends on their basic classification: vowels (weight 12), sonorants (4), voiced (2) and voiceless (1). When the sonorant р (Latin transcription: r) is between two consonants, it becomes a syllable carrier, and therefore its sonority is higher, initially 6. Two adjacent vowels are separated by a fictitious consonant FC.

The sonority of Serbian phonemes is more complex and embraces additional classes: vowels (12), sonorant р (8), sonorants and plosive voiced phonemes (4), plosive voiceless and fricative voiced (3), fricative voiceless and voiced affricates (2), and voiceless affricates (1).

The syllable nuclei in both languages are the five vowels. In Macedonian, a nucleus can be the sonorant *р* appearing within a consonant group (*крст*, *вр-ста*, *пр-вен-ство*) or at the end of the word (*ма-са-кр*). In Serbian language, apart from the sonorant *р* (*тврд*, *црв*, *тр-ка*), the sonorants *л* and *н* can also become syllable nuclei (for example, би-ци-кл, Вл-та-ва, Њу-тн). They are determined by calculating the triplet difference between the sonority of the current phoneme and its left and right neighbours.

Determination of syllable boundaries depends on the monotonically non-decreasing and decreasing sonority. In Macedonian, whenever the sonority of two consonants is non-decreasing, they are split into two adjacent syllables. In Serbian, in the same case both consonants are part of the second syllable.

In Macedonian, the accuracy of the baseline algorithm was rather low, mainly because the suffixes ски, ство and ствен and their inflections, which should remain within one syllable were separated. By adjusting this, we achieved an accuracy of 95.60% evaluated on a corpus of more than 1000 words. However, it affected the syllabification of the nouns: гус-ки, мас-ки, прас-ки, in which ски is not a morpheme.

Based on the sample of more than 3000 syllabified Serbian words, the accuracy of the baseline algorithm was 97.59%. By modifying the sonority of р to 6, the accuracy reached 98.54%, exceeding the rule-based syllabification accuracy.

The approach we proposed is extremely simple and at the same time, very efficient. We intend to further improve it by taking into account the PoS tags for the Macedonian language and the exclusions for Serbian, hoping to reach an accuracy of over 99%.

Keywords: *Macedonian, Serbian, phoneme sonority, syllabification*

References

[1] G. Clements: The sonority cycle and syllable organization, Phonologica,. 63-76, 1988.

[2] M. Mitreska, K. Zdravkova: Syllable and Morpheme Segmentation of Macedonian Language. Proceeding of 46th MIPRO, 1113-1118. IEEE, 2023.

[3] A. Kovač, M. Marković M: A Mixed-principle Rule-based Approach to the Automatic Syllabification of Serbian. Contributions to Contemporary History / Prispevki za Novejšo Zgodovino. 2019.

Danilo Aleksić
Department of Serbian with South Slavic Languages
Faculty of Philology, University of Belgrade
E-mail: danilo.aleksic@fil.bg.ac.rs

New Enhanced Versions of the Program “Ka Minimalnim Parovima” (“Towards Minimal Pairs”)

The short Python program “Ka minimalnim parovima” (“Towards Minimal Pairs”; “KaMP” was chosen as the Serbian abbreviation), first presented in Алексић and Шандрих 2021, was primarily conceived as a tool which teachers of Serbian as a foreign language can use to gather word pairs for pronunciation exercises. KaMP automatically excerpts segmental minimal pairs, e.g. lak ~ luk (*лак ~ лук*), and the word pairs which are related to segmental minimal pairs (i.e. the pairs which would be segmental minimal pairs if prosody were ignored), e.g. gel ~ gen (*гел*

~ *ген*), from a Latin Serbian input corpus encoded in Unicode. In Aleksić and Mrkela 2022, faster variants of KaMP, KaMP 2 and KaMP 2.1 respectively, are presented, both with improved sorting and with a supplementary mode. Now, even faster variants of KaMP are presented, KaMP 2.2 and KaMP 2.3. According to the averages of 20 consecutive measurements per version, KaMP, KaMP 2, KaMP 2.1, KaMP 2.2, and KaMP 2.3 took 122.66, 74, 63.84, 56.51, and 52.74 seconds respectively to excerpt segmental minimal and related pairs which differ from each other by substrings "dž" and "đ" from the corpus POL.xml in Python 3.12.4. The speedup was achieved mainly through the use of classes instead of certain dictionaries in the code. As can be seen, KaMP 2.2 and KaMP 2.3 save noticeable end user’s time when run on large corpora (POL.xml numbers around 117.900.900 words [Алексић and Шандрих 2021: 575]).

A piece of information from Python 2024 (“Attribute lookup speed can be significantly improved . . .”) led to the expectation that the version of KaMP in which classes with the class variable __slots__ are used (KaMP 2.2) will be faster than the version of KaMP in which classes without that class variable are used (KaMP 2.3), but the given measurement results defy the expectation. Since this relative slowness of the slotted classes is somewhat suspicious, it is necessary to show the definitions with the class variable __slots__ which were used in KaMP 2.2:

```
class Dictionary:
    tokens_ = tokenization()
    __slots__ = tokens_

    def __init__(self):
        for word in self.tokens_:
            setattr(self, word, word.casefold())

class Dictionary2:
    __slots__ = tuple_2[1]

    def __init__(self):
        for tuple_ in tuple_2[0]:
            if not hasattr(self, tuple_[0]):
                setattr(self, tuple_[0], {})
            getattr(self, tuple_[0])[
                tuple_[1]] = tuple_[2]

class Dictionary3:
    __slots__ = tuple_2[1]

    def __init__(self):
        for tuple_ in tuple_2[0]:
            if not hasattr(self, tuple_[0]):
                setattr(self, tuple_[0], {})
            getattr(self, tuple_[0])[
                tuple_[1]] = tuple_[2]

def excerpt(letter):
    return (word
            for word in dictionary_from_corpus.__slots__
            if letter in getattr(dictionary_from_corpus, word))
```

The given measurement results can be of use to programmers, because they show that (1) classes can be faster than dictionaries and that (2) unslotted classes can be faster than slotted classes (provided that slots were used properly, i.e. as efficiently as possible, in KaMP 2.2) in Python 3.12.4.

Keywords: *minimal pairs, phonetics, phonology, natural language processing, corpus linguistics, Python*

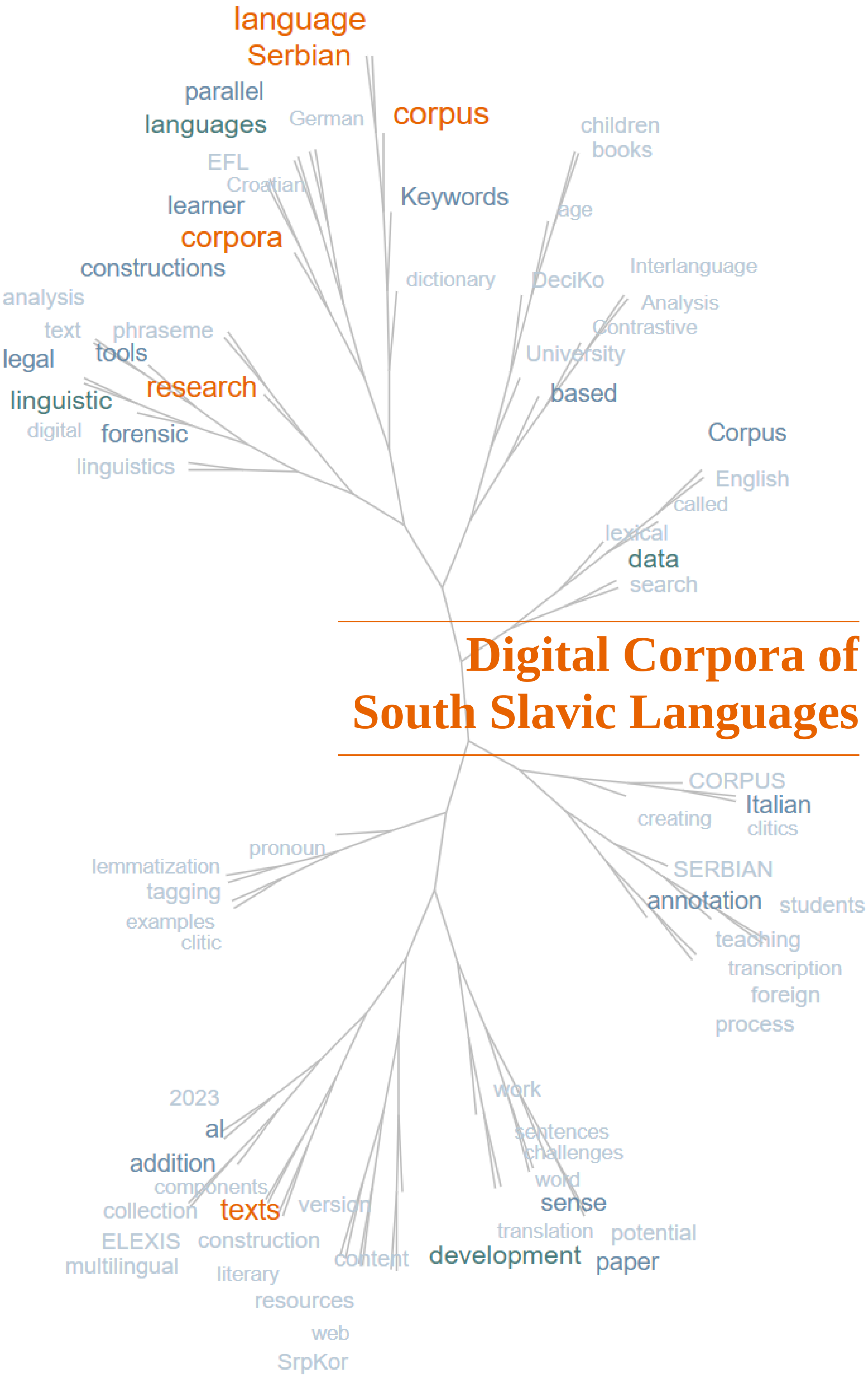
References

[1] Aleksić, Danilo, and Lazar Mrkela. 2022. "The Enhanced Versions of the Program "Ka Minimalnim Parovima" (Towards Minimal Pairs)." *Infotheca* 22 (1): 7–31. https://infoteka.bg.ac.rs/ojs/index.php/Infoteka/article/view/2022.22.1.1_en.

[2] Python 2024: Python 3.12.5 documentation. Accessed August 22, 2024. https://docs.python.org/3/.

[3] Алексић, Данило, and Бранислава Шандрих. 2021. "Аутоматска ексцерпција парова речи за учење изговора у настави српског као страног језика." *Српски језик* 26 (1): 567–584. https://doi.org/10.18485/sj.2021.26.1.3.

Digital Corpora of South Slavic Languages



Jelena Redli
Faculty of Philosophy, University of Novi Sad
E-mail: redli@ff.uns.ac.rs

The Importance of a Digital Corpus and Linguistic Tools in the Linguistic Analysis of Forensic Texts in the Serbian Language

Forensic linguistics (FL) is an interdisciplinary field that combines linguistic, legal, and digital technologies to enhance the understanding of linguistic phenomena in legal and criminal contexts. For nearly a century, FL has been applied in various areas such as author identification, copyright infringement detection, and the discovery of potential criminal activities. The existence of appropriate corpora, used to train systems during their development, is crucial for any FL activity. While the use of corpora in forensic research is becoming an indispensable tool globally, the Serbian language remains significantly under-researched in this field due to the lack of a specialized corpus of forensic texts.

This paper explores the potential for creating and applying a forensic text corpus in the Serbian language, with an emphasis on how such a corpus could revolutionize text authenticity analysis, author identification, and the resolution of legal disputes through linguistic evidence. The paper proposes a methodology for forming a corpus that would include various types of forensic texts, such as police reports, legal documents, threatening messages, suicide notes, and other documents relevant to the judicial context. Strategies for applying advanced linguistic tools for efficient linguistic analysis will also be presented, including software tools for automatic text analysis and opportunities for in-depth analysis of lexical frequency, syntax, stylistic markers, linguistic hedges, genderlects, and more.

In addition to its theoretical contribution, this paper aims to practically demonstrate how the implementation of such a corpus could improve the accuracy of forensic analyses in the Serbian criminal and legal context. The paper also discusses potential challenges in developing the corpus, including technical, legal, and regulatory aspects, as well as possible obstacles in collecting and processing sensitive data. Particular emphasis is placed on the necessity of adhering to ethical and legal guidelines during the development of these resources.

Finally, the paper provides practical recommendations for future research and implementation in practice and offers a foundation for the further development of forensic linguistics in Serbia. It also highlights the importance of interdisciplinary collaboration between linguists, legal professionals, law enforcement, and IT experts to develop large linguistic resources that can serve as key tools in forensic research.

Keywords: *forensic linguistics, Serbian language, legal language, digital corpus, linguistic tools, forensic text analysis, author identification*

References

[1] Blackwell, S. (2009). Why forensic linguistics needs corpus linguistics. *Comparative Legilinguistics, 1*, 5–19.

[2] Chaski, C. (2005). Who's at the keyboard? Authorship attribution in digital evidence investigations. International Journal of Digital Evidence 4(1), 1–14.

[3] Cotterill, J. (2010). How to use corpus linguistics in forensic linguistics. U A. O'Keefe i M. McCarthy (ur.), The Routledge Handbook of Forensic Linguistics, 578–590.

[4] Coulthard, M. (1994). On the use of corpora in the analysis of forensic texts. *International Journal of Speech, Language and the Law, 1*(1), 27–43.

[5] Goźdź-Roszkowski, S. (2021). Corpus linguistics in legal discourse. *International Journal for the Semiotics of Law-Revue internationale de Sémiotique juridique*, *34*(5), 1515–1540.

[6] Hercigonja-Szekeres, M., Sikirica, N., & Popović, I. (2012). Statistička analiza tekstnih podataka. *In medias res: časopis filozofije medija*, *1*(1), 79–91.

[7] Lalić, A. (2024). Elektronski korpus sms poruka na bosanskom jeziku (Halid Bulić, Elma Durmišević, Azra Hodžić-Čavkić, Enisa Bajraktarević, Azra Ahmetspahić-Peljto, Belmin Šabić, Sarajevski korpus SMS poruka na bosanskom jeziku, Univerzitet u Sarajevu–Filozofski fakultet, Sarajevo, 2023). *Društvene i humanističke studije*, *9*(1 (25)), 1187–1190.

[8] Longhi, J. (2021). Using digital humanities and linguistics to help with terrorism investigations. *Forensic Science International*, *318*, 110564.

[9] Vitas, D., Krstev, C., Obradovic, I., Popovic, L., & Pavlovic-Lazetic, G. (2003, November). An overview of resources and basic tools for processing of Serbian written texts. In *Proc. of the Workshop on Balkan Language Resources, 1st Balkan Conference in Informatics*.

[10] Vitas, D., Krstev, C., Obradović, I., Popović, L., & Pavlović-Lažetić, G. (2003, November). Processing Serbian written texts: An overview of resources and basic tools. In *Workshop on Balkan Language Resources and Tools* (Vol. 21, pp. 97–104).

[11] Wright, D. (2017). Using word n-grams to identify authors and idiolects: A corpus approach to a forensic linguistic problem. *International journal of corpus linguistics*, *22*(2), 212–241.

[12] Wright, D. (2020). Corpus approaches to forensic linguistics: Applying corpus data and techniques in forensic contexts. In *The Routledge handbook of forensic linguistics*, 611–627. Routledge.

Milica Dinić Marinković

Univerzitet u Beogradu – Filološki fakultet, Katedra za opštu lingvistiku
E-mail: milica.dinic.marinkovic@fil.bg.ac.rs

Milena Oparnica

Istraživačko-razvojni institut za veštačku inteligenciju Srbije
E-mail: milena.oparnica@ivi.ac.rs

DeciKo – The Corpus of Books for Early Childhood (Age 3 - 7). Current Status and Challenges

The impact of (not) engaging in mediated reading from an early age on children's language and cognitive development is strongly supported by numerous empirical studies (for a review, see Nation et al., 2022). However, the causal relationship between exposing children to the language of books and enhancing their communicative competence remains underexplored. This gap stems from the insufficient attention given to the linguistic features of the input found in books for early childhood and preschool-aged children. To address this issue, a robust and representative collection of machine-readable texts, or a corpus, is essential for conducting systematic linguistic analyses.

With this objective, the project to create the *DeciKo* corpus of books for the children of early age (3 to 7) was initiated in 2022 at the Department of General Linguistics and the Center for Applied Linguistics of the Faculty of Philology, University of Belgrade.

The *DeciKo* Corpus is designed as a dynamic, specialized corpus that is updated annually. It is currently structured based on three extralinguistic criteria: (1) genre, (2) the language of the original, and (3) the century of the first edition of the samples. It is available upon request to authors in TEI XML format. Currently, it contains over 250 children's books. These books have been converted into text format using digital tools, with errors subsequently corrected manually.

In this presentation, we will showcase the current state of the *DeciKo* corpus and discuss the challenges and issues we have encountered during its construction.

A major challenge in constructing the *DeciKo* corpus is the inability to categorize the samples according to the age group they are intended for. Different publishers categorize (even the same) books by age groups in various ways, and the results of pilot studies on the linguistic features of the samples reveal significant deviations from the established patterns of first language acquisition. The fact that this is not merely a local issue is underscored by the existing corpora of books for early childhood and preschool-aged children, which include genre-based, but not age-based divisions of samples. Given that *DeciKo* is designed to be a corpus with full potential for research in first language acquisition and literacy acquisition, it is necessary to find a way to overcome this problem.

Keywords: *DeciKo, corpus of children's books, sampling problem, first language acquisition*

References

[1] Nation, K., Dawson, N., & Hsiao, Y. (2022). Book Language and Its Implications for Children's Language, Literacy, and Development. *Current Directions in Psychological Science, 31*, 375–380.

Dušanka Vujović
Hankuk University of Foreign Studies in Seoul
Faculty of East European & Balkan Studies, Department of South Slavic Studies
E-mail: dusanka.vujovic@ff.uns.ac.rs

Branko Milosavljević
Univerzitet u Novom Sadu - Fakultet tehničkih nauka
E-mail: mbranko@uns.ac.rs

Corpus Srpko as a Technological Basis for the Creation of Dictionary of Contemporary Serbian Language

The digitisation of language resources facilitates easy and quick access to language data, linguistic research, as well as the analysis of linguistic phenomena. It is also crucial for contemporary lexicographic projects because it provides plenty of carefully selected lexical material, and makes it easier to search and find examples, as well as their adding into the dictionary. The electronic language corpus *Srpko* is the basis for the creation of a new Dictionary of the contemporary Serbian language of Matica srpska. The electronic language corpus enables a simple and quick search of textual data and its incorporation into the content of the dictionary text. The paper discusses the principles, and the conceptual and technical solutions related to the organization of the corpus. It shows the structure of the corpus, which consists of the literary, newspaper, scientific, administrative, and conversational subcorpora. Additionally, it explains the method of collecting data and forming a database for each subcorpus, considering the different formats of data, such as written digitized and non-digitized texts, audio, and video. With the aim of faster and easier search of the corpus, but also its control, various functional searches and requirements for filtering results have been incorporated into the program, which helped reduce the noise, i.e., redundant information that congests search results. The function of obtaining various reports based on specified filters is also built into the program. These can be ready-made reports obtained based on predetermined parameters or dynamic reports that can be created as needed based on specified filters. This approach to the corpus lays the foundation for systematically organizing and presenting the lexical database, leading to the creation of a dictionary.

Keywords: *Serbian language, Serbian language corpus, digitalization, dictionary*

Kristina Ilić
University of Belgrade, Faculty of Philology
E-mail: kristina.ilic997@gmail.com

The Importance of Parallel Corpora for the Research of Phraseme Constructions in the German and Serbian Languages

Parallel corpora are of great importance for contrastive studies of phraseme constructions in different languages. According to Dobrovoljski (e.g. 2011: 114), phraseme constructions are defined as lexically partially specific pairs of forms and meanings (= constructions), whose semantics derives not only from the lexical meaning of their components and the manner of their syntactic connection but also from the construction as a whole. Phraseme constructions are idiomatic combinations of words that consist of certain lexical elements, so-called anchors, and a certain number of empty spaces, the so-called slots. These slots are filled with the so-called fillers when using the structure (Dobrovol’skij, 2011, p. 114). As confirmed by Ďurčo, modern monolingual and multilingual corpora provide new tools for comparing data from large corpora of two languages, which consist of independent texts (Ďurčo, 2018, p. 117). Constructional grammar is a young branch of linguistics, and the research of phraseme constructions with the help of corpora, especially parallel corpora in the case of contrastive research, is developing increasingly. This is especially the case if the goal of the corpus search is to find translation equivalents of constructions in the source language in parallel texts in the target language, where these translation equivalents in the target language are often constructions themselves, which contributes to the research of phraseme constructions in the target language as well. At the moment, it can be said that the parallel corpora of the German and Serbian languages are less represented, in contrast to the parallel corpora of German and other major world languages, such as *COMBIDIGILEX* or *EuReCo*, or the parallel corpora of other major world languages and the Serbian language, such as *SrpEngKor*, *SrpFranKor*, *Evroteka* or *ParCoLab*. This talk aims to present the currently available resources in the form of parallel corpora of the German and Serbian languages, such as *SrpNemKor* within the tool *Bibliša*, corpora that are part of *Sketch Engine* or the corpus *InterCorp* within the Czech *National Corpus*. In addition, it will show the possible ways of their usage in the contrastive research of phraseme constructions on concrete examples, point out what else is needed for such research, and suggest the layout, scope, search possibilities, etc. of an optimal resource. This research is carried out as part of the COST action PhraConRep "A Multilingual Repository of Phraseme Constructions in Central and Eastern European Languages", which aims, among other things, to investigate the phraseme constructions of German and of the Serbian language contrastively, whereby the use of different tools and parallel corpora, presented in this research, can play a crucial role in reaching that goal.

Keywords: *Construction grammar, corpus linguistics, phraseme constructions, Serbian language, German language*

References

[1] Dobrovol'skij, D. (2011). Phraseologie und Konstruktionsgrammatik. In A. Lasch. & A. Ziem (eds.), *Konstruktionsgrammatik III. Aktuelle Fragen und Lösungsansä*tze (pp. 110–130). Tübingen: Stauffenburg.

[2] Ďurčo, P. (2018). Vom Nutzen der vergleichbaren Korpora bei der kontrastiven lexikographischen Erfassung von Mehrworteinheiten. In V. Jesenšek & M. Enčeva, (eds.), *Wörterbuchstrukturen zwischen Theorie und Praxis* (pp. 107–118). Berlin, Boston: De Gruyter.

Nevena Ceković
Univerzitet u Beogradu – Filološki fakultet
E-mail: n.cekovic@fil.bg.ac.rs

Revision of Errors in the Italserb Learner Corpus

Orthographic transcription represents a crucial step in creating a usable spoken corpus. However, the complexity of this methodological process produces, as expected, numerous potential errors. Especially, when dealing with a learner or corpus of second language, and when creating such a corpus is a pioneering effort in the field of applied corpus linguistics in Serbia. The homogeneity of transcription records is also a particular methodological challenge, considering that multiple researchers are involved in the process.

The study provides an insight into the revision process of the ITALSERB (*ITALiano dei SERBofoni*) corpus of Serbian university students of Italian as a foreign language, which has been realized since 2010 at the Department of Italian Studies (Faculty of Philology, University of Belgrade). This large corpus consists of approximately 25 hours of audio recordings of asymmetric interactions involving over 170 subjects, with varying language proficiency levels (from A2 to C1). The corpus is currently under revision of the transcription process and morphosyntactic annotation of the collected linguistic data.

The study specifically examines the corrections of errors occurring as a result of the transcription process, as opposed to those typical for leaner corpora, which reflect characteristic stages in the development of competence and therefore provide valuable insights into the specific features of interlanguage. The goal of the study is to provide guidelines for good practice in constituting (transcribing and annotating) a corpus of second language through the illustration of selected examples of transcribed excerpts from the corpus for revision purposes.

Keywords: *ITALSERB, learner corpus, L2 corpus, Italian as a foreign language, Serbian students, transcription, revision*

Jelena Marković
University of East Sarajevo
E-mail: jelena.markovic@ff.ues.rs.ba

Serbian EFL Learner Corpora Within Contrastive Interlanguage Analysis

The development of Contrastive Interlanguage Analysis, induced by the necessity to innovate the contrastive methodology in applied linguistics, has been directly influenced by the appearance and forming of learner corpora, like the International Corpus of Learner English (ICLE). Nowadays, Contrastive Interlanguage Analysis is seen as a renowned and compulsory methodology in the discipline of Learner Corpus Research.

A considerable advance regarding Contrastive Interlanguage Analysis and learner corpora has been noticeable in Serbian-speaking areas. Having said that, we bear in mind that the successful collection of the EFL learner corpus called *ICLE-SE* (*International Corpus of Learner English – Serbian*), which is a subcorpus of the aforementioned ICLE corpus, has been of great significance. Besides this, it is also important to mention the complementary learner corpus called *KorSAng* (Corpus of English-Studies Students), which incorporates bidirectional student translations and argumentative essays in native Serbian. The two learner corpora are analysed from the points of view of the relevant metadata and the annotation.

Having introduced the topic, we aim to report about the Contrastive Interlanguage Analysis research based on the two Serbian EFL learner corpora published so far, offering a critical overview followed by a discussion about the directions of possible development. More precisely, we discuss the research on metadiscourse features, discourse connectors and the phenomenon of lexical vagueness, accompanied by an analysis of the word frequency lists of the EFL learner corpora and their referential native corpora. The presentation then focuses on the scientific potential of the Serbian EFL learner corpora which has not been recognised adequately so far. Finally, we aim to offer suggestions for the future popularisation of learner corpora in Serbian-speaking areas, accompanied by an analysis of potentially widening the two EFL learner corpora.

Keywords: *learner corpus, ICLE-SE, KorSAng, Contrastive Interlanguage Analysis, metadata, variables, discourse*

References

[1] Ädel, A. (2006). *Metadiscourse in L1 and L2 English* (Studies in Corpus Linguistics ed., Vol. 24). Amsterdam/Philadelphia: John Benjamins.

[2] Deshors, S. C., Götz, S., & Laporte, S. (2016). Linguistic innovations in EFL ansd ESL: Rethinking the linguistic creativity of non-native English speakers. *International Journal of Learner Corpus Research*, *2*(2), 131–150.

[3] Gilquin, G. (2008). Combining contrastive and interlanguage analysis to apprehend transfer: detection, explanation, evaluation. In G. Guilquin, S. Papp, & M. B. Díez-Bedmar (Eds.), *Linking up contrastive and learner corpus research* (pp. 3–33). Amsterdam and New York: Rodopi.

[4] Granger, S. (2012). How to use second and foreign language learner corpora. In A. Mackey, & S. M. Gass (Eds.), *Research methods in Second Language Acquisition: A Practical Guide* (pp. 7–29). Malden: Blackwell.

[5] Granger, S. (2015). Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research, 1*(1), 7–24.

[6] Lee, D. Y., & Chen, S. X. (2009). Making a bigger deal of the smaller words: Function words and other key items in research writing by Chinese learners. *Journal of Second Language Writing, 18*, 281–296.

[7] Marković, J. (2019). “It is a thing that gives you…“: The lexeme thing(s) as the Serbian EFL ‘teddy bear’. *Komunikacija i kultura online, 10*, 19–37.

[8] Marković, J. (2021). *I*, *you*, and *we* in Serbian EFL Argumentative Writing from the Essay Title Perspective. *Folia Linguistica et Litteraria, 36*, 271–290. doi: 10.31902/fll.36.2021.16

[9] Nesselhauf, N. (2005). *Collocations in a Learner Corpus* (Studies in Corpus Linguistics ed., Vol. 14). Amsterdam/Philadelphia: John Benjamins.

[10] Šućur, S. (2019). Distribucija frazalnih glagola u pisanju na engleskom kao stranom kod srbofonih govornika. *Komunikacija i kultura online, 10*, 120–143.

[11] Tomović, N., & Marković, J. (2020). The Status of English in Serbia. In S. Granger, M. Dupont, F. Meunier, H. Naets, & M. Paquot (Eds.), *The International Corpus of Learner English. Version 3.* (236–242). Louvain-la-Neuve: Presses universitaires de Louvain.

[12] Марковић, Ј. (2017). Лични метадискурс у писању код неизворних и изворних говорника енглеског језика. *Филолог, VII* (15), 44–60.

[13] Марковић, Ј. (2018). Употребе глагола *make* у писању на енглеском језику као страном код изворних говорника српског језика (корпуснолингвистичка анализа). *Зборник Матице српске за филологију и лигвистику, LXI* (1), 165–180.

[14] Марковић, Ј. (2019). *Кроз призму контрастивне анализе међујезика.* Пале: Филозофски факултет.

[15] Марковић, Ј. (2020). Концесивни конектори *though* и *however* у писању на енглеском језику код изворних и неизворних говорника. *Филолог*, 21, 13–35.

[16] Марковић, Ј. и Станковић, Р. (2021). *Ја/ти/ми/ви* у дискурсној компетенцији у светлу контрастивне анализе међујезика. *Методички видици*, 12, 95–119.

[17] Радоња, М. и Шућур, С. (2021). О Корпусу студената англистике (КорСАнг) и могућностима његове софтверске експлоатације. *Infotheca – Journal for Digital Humanities, 21(1)*, 37–58.

Saša Moderc
Univerzitet u Beogradu – Filološki fakultet
E-mail: moderc.sasa@gmail.com

Italian Clitics and Tagging: User Experiences with the Serbitacor3 Corpus

Italian possesses eleven clitics serving various grammatical functions, including personal pronouns, adverbial particles, passive markers, and impersonal structure markers. These clitics can also be used pleonastically and are integral components of pro-complement verbs (e.g., *farcela* ‘to manage’, derived from the basic verb *fare* ‘to do’). They can occupy enclitic, proclitic, and mesoclitic positions relative to the verb, and can form groups of two or three. Such groups can occupy a single position or split, with one clitic in a proclitic and the other in an enclitic position. The multifunctional nature of most Italian clitics means that determining their function relies on syntactic, semantic, and pragmatic factors, requiring advanced linguistic competence. Such competence is typically possessed by translators of literary works, upon whose texts the Serbian-Italian parallel corpus (SerbItaCor3) is based. For instance, the clitic sequence *lo si è visto* occurs three times in this corpus and is correctly translated into Serbian in all instances. However, while the clitic *lo* serves as a personal pronoun in one case ('they saw **him'**), and as a demonstrative pronoun in two other cases ('as we saw', 'we already saw that'), it is consistently tagged as a personal pronoun in SerbItaCor3, following the instructions contained in the Tree-Tagger tool developed by Achim Stein. This study identifies other tagging inaccuracies, particularly concerning the clitic *si*, which is sometimes incorrectly labeled as a personal pronoun (as in *si guarda*, where *si* could be a reflexive pronoun, a passive marker, or an impersonal marker). Despite these inaccuracies, the translations are accurate, highlighting the limitations of the automated tagging process. These inaccuracies reduce the reliability of the linguistic data in the corpus and limit its applicability in language teaching. Nevertheless, the accurate translations allow for the identification of the correct functions of clitics, providing valuable data for improving the tagger by supplementing the Tree-Tagger rules with information about verb valency, arguments, and adverbial phrases. This paper presents examples of inaccurate clitic tagging and the corresponding corrective Serbian translations. It also outlines linguistic parameters that can be used to develop instructions for improving existing taggers or creating a new Italian language tagger.

Keywords: *clitics, Italian, Serbian, corpus, tagging, tagger improvements*

References

[1] Bentley, D. (2006). Split Intransitivity in Italian. Berlin-New York. Mouton de Gruyter.

[2] Dell'Orletta F. (2009). Ensemble system for Part-of-Speech tagging. In: Proceedings of EVALITA 2009 – Evaluation of NLP and Speech Tools for Italian. Reggio Emilia, Italy.

[3] Moderc S., Stanković R., Tomašević A., Škorić M. (2023). An italian-serbian sentence aligned parallel literary corpus. Review of the National Center for Digitization, 43.

[4] Moderc, S. (2021). I clitici italiani. Usi, ambiguità, interpretazioni. Volume primo: il sistema dei clitici. Beograd. Filološki fakultet.

[5] Moderc, S. (2021). I clitici italiani. Usi, ambiguità, interpretazioni. Volume secondo: i nessi di clitici. Beograd. Filološki fakultet.

[6] Moderc, S. (2015). Gramatika italijanskog jezika. Morfologija s elementima sintakse. Beograd. Luna crescens.

[7] Renzi, L. (1988). Grande grammatica italiana di consultazione. Vol. 1. Bologna. Il Mulino.

[8] Russi C. (2008). Italian Clitics. An Empirical Study. Berlin-New York. Mouton de Gruyter.

[9] Salvi, G., Vanelli, L. (2004). Nuova grammatica italiana. Bologna. Il Mulino.

[10] Schmid, H. (2013). *Probabilistic part-of speech tagging using decision trees*, In New methods in language processing, 5.

[11] Serianni, L. (1989). Grammatica italiana. Italiano comune e lingua letteraria. Torino. Utet.

[12] Tamburini F. (2009). PoS-tagging Italian texts with CORISTagger. In: EVALITA 2009. Workshop on Evaluation of NLP and Speech Tools for Italian. vol. 1. Reggio Emilia, Italy, December 2009.

[13] Tamburini, F. (2000). Annotazione grammaticale e lemmatizzazione di corpora in italiano. In: R. Rossini, Linguistica e informatica: multimedialità, corpora e percorsi di apprendimento. Bulzoni, Roma, 2000.

Olja Perišić
Univerzitet u Torinu
E-mail: olja.perisic@unito.it

Corpora for Learning Serbian as a Foreign Language in the Era of Artificial Intelligence

The introduction of corpora in foreign language teaching is associated with the 1990s and an approach known as DDL (Data-Driven Learning) in teaching English as a foreign language. While researchers active in this field are re-examining the future of corpus tools in the era of artificial intelligence (Crosthwaite and Baisa 2023, Flowerdew 2024), the Serbian language remains in its infancy, at least regarding the practical use of corpora in teaching. Namely, there is a significant gap between, on one hand, the research activities of IT experts (especially those gathered around the JeRTeh association) who follow global trends and actively work on developing corpus tools and, on the other hand, the teaching staff who could use these tools in working with students. In addition to the necessary affinity for this type of teaching, one of the causes of this discrepancy is the initial investment in mastering corpus search skills, as well as adopting a theoretical and methodological approach that enables classroom work. Experience has shown that students of Serbian as a foreign language relatively quickly master the skill of using corpora and are generally creative in independent research (Perišić 2023). It has also been shown that, thanks to corpora, students develop a sense of language, an inclination towards research work, and motivation for learning Serbian from the initial level. In the era of accelerated development of artificial intelligence and the growing trend of declining numbers of philology students, not only for our language, we will highlight the advantages of corpus research in foreign language learning, with an emphasis on Serbian as a foreign language, and point out the challenges but also the advantages of corpus tools compared to generative language models.

Keywords: *corpora, Serbian as a foreign language, Data Driven Learning, JeRTeh, artificial intelligence*

References

[1] Crosthwaite Peter, Baisa Vit (2023). *Generative AI and the end of corpus-assisted data-driven learning? Not so fast!*, “Applied Corpus Linguistics 3”, https://doi.org/10.1016/j.acorp.2023.100066

[2] Flowerdew John (2024), Data-driven learning: From Collins Cobuild Dictionary to ChatGPT, *Language Teaching*, 1–18.

[3] Perišić Olja (2023), *Il corpus per imparare il serbo. Il futuro dell’apprendimento linguistico*, Alessandria: Edizioni dell’Orso.

Simon Krek
"Jožef Stefan" Institute, Artificial Intelligence Laboratory
University of Ljubljana, Centre for Language Resources and Technologies
E-mail: simon.krek@ijs.si

Carole Tiberius
Faculty of Humanities, Leiden University
Centre for Linguistics
E-mail: carole.tiberius@ivdnt.org

Jaka Čibej
Faculty of Arts, University of Ljubljana
Centre for Language Resources and Technologies, University of Ljubljana
E-mail: jaka.cibej@ff.uni-lj.si

Ana Ostroški Anić
Department of General Linguistics
Institute for the Croatian Language
E-mail: aostrosk@ihjj.hr

Ranka Stanković
University of Belgrade, Faculty of Mining and Geology
Chair for Mathematics and Informatics
E-mail: ranka@rgf.rs

Extension of the Elexis-WSD Parallel Sense-Annotated Corpus with South Slavic Languages: Challenges, Results, and Plans

The open-source ELEXIS-WSD Parallel Sense-Annotated Corpus (Martelli et al. 2021; Martelli et al. 2023) was developed within the ELEXIS project and in version 1.1 contains 2,024 sentences for each of 10 languages: Bulgarian, Danish, English, Spanish, Estonian, Hungarian, Italian, Dutch, Portuguese, and Slovene. Within the sentences, each content word (nouns, adjectives, verbs, and adverbs) has been assigned a corresponding sense from one of the 10 sense inventories containing definitions.

Within the context of Task 2.2 of the UniDive COST Action (CA21167), the corpus will be extended with several new languages, including at least two South Slavic languages: Croatian and Serbian. The process of extending the corpus involves several different stages from translation to tokenization, lemmatization, and POS-tagging, to named entity and multiword expression annotation, and finally, word-sense disambiguation. In this paper, we present the challenges encountered in these different phases when extending the corpus with Serbian and Croatian, as well as the process of adding additional annotation layers for the Slovene part of the corpus.

The Serbian addition to the ELEXIS multilingual annotated corpus started with the automatic translation of sentences, followed by manual postediting. The tokenization, POS tagging, lemmatization, named entity recognition, and linking were also manually checked (Krstev et al. 2024). The recognition of multiword expressions (MWEs) will be elaborated. The first steps and challenges in building the Serbian sense inventory will be discussed, and some results concerning MWEs and NEs will be analyzed. Once completed, the ELEXIS-WSD-SR corpus will be the first sense annotated corpus using the Serbian WordNet (SrpWN).

The Croatian addition of the ELEXIS corpus started with the same procedure of automatic translation and manual validation of the sentences. However, additional work was needed to prepare the sense inventory for the Croatian language that is to be used for word-sense disambiguation. The resource used as the base for the task is a dictionary in XML format, the data categories of which are not structured in accordance with the existing lexical or lexicographical data models. Therefore, the data was first parsed into separate data categories, followed by extensive verification that different labels (e.g., grammatical and usage labels) were recognized according to their original purpose. The extension of the resource and its conversion into an open sense repository is further discussed.

Like the previous version, the extended corpus and its corresponding sense inventories will be made available at the CLARIN.SI repository under CC-BY-SA 4.0.

Acknowledgment: *This research was supported by the COST ACTION CA21167 - Universality, Diversity, and Idiosyncrasy in Language Technology (UniDive). The authors also acknowledge the financial support from the Slovenian Research and Innovation Agency (research core funding No. P6-0411 – Language Resources and Technologies for Slovene).*

Keywords: *semantic annotation, parallel corpus, senses, South Slavic languages, Slovene, Croatian, Serbian*

References

[1] Martelli, Federico, Roberto Navigli, Simon Krek, Jelena Kallas, Polona Gantar, Veronika Lipp, Tamás Váradi, András Győrffy, and László Simon. "Designing the ELEXIS parallel sense-annotated dataset in 10 European languages." (2021): 377-395.

[2] Cvetana Krstev, Ranka Stanković, Aleksandra Marković, Teodora Mihajlov “Towards the semantic annotation of SR-ELEXIS corpus: Insights into Multiword Expressions and Named Entities, Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD 2024), pp. 74-84, SIGLEX, ACL, UniDive CA21167. eds.: Bhatia, Archna et al. virtual presentation, May 25, 2024. https://aclanthology.org/2024.mwe-1.15.pdf

[3] Martelli, Federico, Roberto Navigli, Simon Krek, Jelena Kallas, Polona Gantar, Svetla Koeva, Sanni Nimb, Bolette Sandford Pedersen, Sussi Olsen, Margit Langemets, Kristina Koppel, Tiiu Üksik, Kaja Dobrovoljc, Rafel Ureña-Ruiz, José-Luis Sancho-Sánchez, Veronika Lipp, Tamás Váradi, András Győrffy, László Simon, Valeria Quochi, Monica Monachini, Francesca Frontini, Carole Tiberius, Rob Tempelaars, Rute Costa, Ana Salgado, Jaka Čibej, Tina Munda, Iztok Kosem, Rebeka Roblek, Urška Kamenšek, Petra Zaranšek, Karolina Zgaga, Primož Ponikvar, Luka Terčon, Jonas Jensen, Ida Flörke, Henrik Lorentzen, Thomas Troelsgård, Diana Blagoeva, Dimitar Hristov, Sia Kolkovska, 2023: Parallel sense-annotated corpus ELEXIS-WSD 1.1, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1842.

Duško Vitas, Ranka Stanković, Cvetana Krstev
Society for language resources and technologies JeRTeh
E-mail: {vitas|cvetana|ranka}@jerteh.rs

The Many Faces of SrpKor

The acronym SrpKor denotes a family of electronic corpora of the modern Serbian language, the construction of which began at the end of the seventies of the last century, and which became more widely visible to the interested research community with the publication of its first version on the web in 2002. In this long period, especially before the emergence of useful textual resources on the web, corpus development consisted of the collection and processing of material as well as the development of corpus processing methods. Namely, an electronic corpus is not only a collection of texts in digital form (as, for example, it is stated in (Dobrić 2012)), but includes several components that will make such a collection useful in linguistic and other research. These components, in addition to the texts themselves, constitute, above all, software support for the organization and exploitation of the collection of texts and means for different levels of annotation of the texts that will be found in the corpus (Ви-тас 2023).

SrpKor, taking into account these components, underwent various metamorphoses during its construction, which provide a picture of the evolution of software support for the construction and exploitation of corpora, as well as the development of annotation systems at different levels (meta-data, morphological marking, lemmatization, named entities, etc.).

Extremely modest conditions (compared to other environments, both in the number of researchers involved in the construction of the corpus, allocated financial resources from different sources, and available equipment) imposed a strategy of gradual development of the corpus, which implied that new versions of the corpus would rely on material prepared and used in those versions that preceded it.

The paper will illustrate the evolution in the development of SrpKor from its first version until today, following the influx of different resources used in the construction of individual versions, as well as the changes in dimensions and text annotation system. The structure of the individual versions of the corpus, their dimensions, the period covered, and the level of annotation will be described in particular.

The basic ideas when conceiving the corpus are first presented in (Vitas, Popović 2023), and then in (Utvić 2013), where numerous details for the 2013 version of SrKor are described. Corpus interactions with dictionaries are discussed in (Krstev Vitas 2005), (Vitas, Krstev 2012).

It is important to note that texts in the Serbian language from parallelized corpora that were created at the same time as SrpKor were also included in SrpKor. In this way, the influence of web content on the composition of the corpus was partially compensated. On the other hand, such texts, which are, as a rule, extremely significant in the cultural sense, not only are not present in the material from the web but are not even included in traditional lexicographic corpora. They consist of selected scientific, literary, philosophical, anthropological, historical, and similar texts taken from reputable editions.

Further work on the development of this corpus will include, on the one hand, the enrichment of metadata, the addition of annotations and the introduction of new content. Enrichment of metadata will enable the creation of subcorpus according to different dimensions: by pronunciation, period, and domain, in addition to the ones available so far by author, register, and years. Along with the division into sentences and the addition of annotations with named entities everywhere, the plan is to enrich them with grammatical information. The introduction of new content expands the time dimension by preparing novels, travelogues, memoirs, and historical newspapers that are valuable not only from a linguistic but also from a cultural-historical point of view, with the usual addition of (selected) content from the web.

The coupling of the Leximirka lexical base and the SrpKor corpus family is two-way, from the Leximirka interface, direct insight into examples of word use in context or in syntactic patterns is possible (Lazić, Škorić 2020). The system for lemmatization is improved from version to version, in which the Serbian morphological dictionaries play a special role, which, using the Unitex system, ensures the generation of all inflectional forms of words.

Keywords: *SrpKor, corpora, Serbian, lemmatization, Leximirka*

Acknowledgment: *This research was supported by the Science Fund of the Republic of Serbia, #7276, Text Embeddings - Serbian Language Applications - TESLA.*

References

[1] Cvetana Krstev, Duško Vitas, “Corpus and Lexicon - Mutual Incompletness”, in Proceedings of the Corpus Linguistics Conference, 14-17 July 2005, Birmingham, eds. Pernilla Danielsson and Martijn Wagenmakers, ISSN 1747-9398, http://www.corpus.bham.ac.uk/PCLC/, 2005.

[2] Dobrić, Nikola. "Savremeni jezički korpusi na Zapadnom Balkanu–Istorijat, trenutno stanje i budućnost (Language Corpora in the West Balkans–History, Current State and Future Perspective)." Slavistična revija 60 (2012): 677-692.

[3] Душко Витас, Љубомир Поповић, „Конспект за изградњу референтног корпуса српског стандардног језика“, Научни састанак слависта у Вукове дане 31/1 - МСЦ, Београд, 2003, стр. 221 - 227.

[4] Душко Витас, Белешке о ручној и аутоматској обради српског језика, Језик Данас, бр. 22, 2023, Матица Српска, Нови Сад.

[5] Duško Vitas, Cvetana Krstev “Tvorbeni obrasci u elektronskom rečniku srpskog jezika”, Međunarodni komitet slavista. Komisija za tvorbu reči. Međunarodna naučna konferencija Tvorba reči i njeni resursi u slovenskim jezicima (14), pp. 515-525, 2012, Filološki fakultet Univerziteta u Beogradu, Beograd, ISBN 978-86-6153-116-3

[6] Utvić, Miloš. 2013. “Izgradnja referentnog korpusa savremenog srpskog jezika.” PhD diss., Univerzitet u Beogradu, Filoloˇski fakulte. https://nardus.mpn.gov.rs/handle/123456789/4091

[7] Biljana Lazić, Mihailo Škorić. "From DELA Based Dictionary to Leximirka Lexical Database" in Infotheca, Faculty of Philology, University of Belgrade (2020). https://doi.org/10.18485/infotheca.2019.19.2.4

Saša Marjanović
University of Belgrade – Faculty of Philology
E-mail: sasa.marjanovic@fil.bg.ac.rs

Dejan Stosic
University of Toulouse Jean Jaurès & CLLE (UMR 5263 CNRS)
E-mail: dejan.stosic@univ-tlse2.fr

Collections of Texts in Serbian in the Nine-Language Parallel Corpus *ParCoLab*

The multilingual—nine-language—parallel corpus ParCoLab (Stosic et al. 2015) is the result of bilateral cooperation between researchers from the University of Toulouse – Jean Jaurès and the Department of Romance Studies at the Faculty of Philology, University of Belgrade. Recently, collaborators from other scientific research organizations in France, including the University of Poitiers, the University of Strasbourg, and the University of Corsica, have also joined the project. The corpus is stored within the French digital infrastructure HumaNum and can be accessed freely—without creating an account—through a specially designed interface for corpus searches on the website: http://parcolab.univ-tlse2.fr.

The development of the corpus began in 2007 with the alignment of literary texts in Serbian and French. Since 2010, texts in English have also been included (Miletić et al. 2014; Miletić et al. 2017; Marjanović et al. 2018; Stosic et al. 2019; Marjanović et al. 2019). The addition of new languages continued in 2018, when texts in Occitan were incorporated. In 2020, the inclusion of Spanish began, followed by the addition of documents in Poitevin-Saintongeais in 2022 and texts in Alsatian and Corsican in 2023. The latter languages, along with Occitan, are part of the regional languages that contribute to the linguistic heritage of France (Stosic et al. 2024). Finally, in 2024, the development of the Italian subcorpus commenced with the integration of bidirectionally aligned texts from the *SerbItaCor3* corpus (Moderc et al. 2023). In addition, throughout this period, work was also done on expanding the domains and genres of texts (e.g., Terzić et al. 2020). Alongside literary texts, the corpus now includes multilingual aligned transcripts of full-length and animated films, newspaper articles, content from multilingual internet portals, as well as translated religious, biological, and cinematographic texts, sports regulations, political speeches, and more. Corpus texts are classified into appropriate special collections based on their domain and origin. Currently, the aligned and searchable texts in all covered languages reach a total of 50,000,000 words. In this presentation, the collections of texts from the multilingual parallel corpus ParCoLab are introduced, highlighting the Serbian language as both the source and target language. The scope of their representation is outlined, and the potential applications in several philological fields are discussed.

Keywords: *ParCoLab, parallel corpus, Serbian language, French language, English language, Spanish language, Italian language, regional languages in France*

Language Resources for South Slavic Languages

entities
named
base
knowledge
existing
entries
automatically
Wikidata
system
present
Dictionary
Language
types
RAKOPS
developmental
text
work
spatial
domain
plans
specific
questions
corpus
answers
texts
challenges
planning
answering
subcorpus
data
models
resources
developed
language
technologies
Serbian
general
collocations
frequencies
extraction
words
SERBIAN
development
focus
research
linguistics
languages
paper
methods
al
semantic
Keywords
annotation
corpora
football
noun
Spanish
verbs
terms
forms
form
hand
users
dictionary
verb
resource
lexicographic
includes
modules
web
application
module
conjugation

Jelena Lazarević
Univerzitet u Beogradu, Filološki fakultet, doktorand
E-mail: jelazarevic1@gmail.com

Olivera Kitanović
Univerzitet u Beogradu, Rudarsko-geološki fakultet
E-mail: olivera.kitanovic@rgf.bg.ac.rs

Contrastive Analysis of Syntax Patterns in Comparable Football Corpora in Spanish and Serbian Languages

The aim of the paper is to explore collocability as a manner in which lexical units are combined with words from different categories, forming larger units. The research of the semantic and syntactic principles of these combinations of Spanish and Serbian footballing terms was carried out on the comparable football corpora *SrFudKo* and *EsFudko* developed as part of Jelena Lazarevic's doctoral dissertation titled: *Language characteristics of the new media discourse on football: a contrastive analysis of the Serbian and Spanish language corpora.*

The football corpus *SrFudKo* was developed through texts about football from five Serbian web news sites: *B92, Blic, Mondo, Politika,* and *Sport klub*, containing 10,100,553 tokens, of which 8,618,426 words. The corpus of Spanish-language texts on football *EsFudKo*, comes from two Spanish sites: *Marca fútbol* and *Mundo deportivo*, containing 9,106,812 tokens, of which 8,024,164 words. Both corpora to which corpus linguistics methods have been applied for data extraction are located on the platform https://noske.jerteh.rs, and are available to authorized users.

In this paper, the mutual lexical-semantic "attractiveness" of collocations is determined based on frequencies and other measures within the corpora, so that collocations are viewed in the broadest sense of Corpus linguistics - as a series of words or concepts that appear together more often than expected by chance. We will present seven main types of collocations through the following examples: adjective + noun *(fast counter)*, noun + noun *(penalty shootout)*, verb + noun *(to score a goal)*, adverb + adjective *(very talented)*, verbs + prepositional phrase *(play at the stadium)* and verb + adverb (to kick hard). Collocation extraction represents a technique in Computational linguistics that identifies collocations in a text or corpus of texts, using elements similar to data mining, while relying on syntactic patterns and frequencies of occurrence.

In addition to frequencies of occurrence, we also consider other factors, such as semantic closeness and context in both languages. For example, do certain collocations have specific meanings, or are they only used in certain situations? We also consider whether the previously identified collocations are understandable to the general public who do not follow sports and are not versed in the language of football. If a speaker from the general understands them, then the collocations have surpassed their origin in the football domain, becoming part of the public domain.

The contribution of the research also means analyzing the connections between collocations and multi-part terms. Their connection is strong when the multi-part terms contain collocates that have a clear meaning within the domain of football. This helps understand the terminological connection within the language of football, providing insight into typical word combinations and their use, illustrating those that often appear in football corpora of the Serbian and Spanish languages of football.

Keywords: *football, corpora, terminology, collocations, Serbian, Spanish*

Acknowledgment: *This research was supported by the Science Fund of the Republic of Serbia, #7276, Text Embeddings - Serbian Language Applications - TESLA.*

Rada Stijović, Ranka Stanković, Mihailo Škorić,
Society for Language Resources and Technologies JeRTeh
E-mail: stijovicr@yahoo.com, {ranka|mihailo}@jerteh.rs

Dictionary of Modern Serbian Language: RSSJ

The presentation will outline the motivation for creating the Dictionary of the Contemporary Serbian Language (RSSJ), its concept and challenges in implementation, as well as examples of some solutions, both from the lexicographic perspective and from the perspective of the software solution of the data model and components of the integral system. The Society for Language Resources and Technologies JeRTeh accepted the initiative of the association "Gathered around the Language" from the diaspora and directed it towards a solution that would, on the one hand, achieve the Association's desired goals: the creation of a paper version of the dictionary and dictionary database, and, on the other hand, be in accordance with contemporary lexicographic and IT standards.

The goal of the project is to create a dictionary with about 50,000 dictionary entries, which will present the lexicon of the modern Serbian standard language. The material for the dictionary was automatically extracted from the SrpKor2013 and SrpKor2021 electronic corpora, as well as from the LeXimirka lexical database. It contains a lexicon of all styles of standard language (literary, scientific, journalistic, administrative and conversational) used in the last fifty years, with the fact that the literary and artistic style also includes somewhat older works (from the Second World War onwards). The vocabulary is processed using the methods of modern lexicography, and the explained meanings are given in a way that is accessible to a wide readership.

Different data driven methods are used for the development of RSSJ, which, in addition to the SrpKor family and Leximirka, include language models trained within the JeRTeh Society and the TESLA (Text Embeddings - Serbian Language Applications) project. The data model is inspired and largely aligned with the Data Model for Lexicography (DMLex), while the database is implemented in the PostgreSQL system.

We will present various ways of automating article creation that rely on artificial intelligence and corpus linguistics methods, from collocation frequencies, extraction of language patterns, to automatic definition generation and extraction of good examples of usage (GDEX, Good Dictionary Examples). The user interface of the lexicographic web application developed for the purposes of writing the Dictionary links the dictionary entry with the corpora on the platform https://noske.jerteh.rs/. More precisely, for the headword of a dictionary article or one of its components (term, synonym, reference) the lexicographic application forwards the appropriate CQL query and parameters for the type of processing: concordance, collocation or frequency of occurrence of forms and terms. Finally, the vocabulary development environment and usage modalities will be presented.

Keywords: *dictionary, Serbian, lexicography, lexicographic base, corpus*

Milena Milinković
Institute of Architecture and Urban & Spatial Planning of Serbia
E-mail: millena.milinkovic@gmail.com

Milica Ikonić Nešić
University of Belgrade - Faculty of Philology
E-mail: milica.ikonic.nesic.fil@gmail.com

Named Entities in the Digital Corpus of Spatial Plans/Planning

This work will present the subcorpus of the exemplary domain-specific corpus of spatial planning, as well as the challenges that occurred during its formation and annotation due to its atypical content. This subcorpus includes six planning documents of different spatial scope, where the focus is on the textual parts of one Regional Spatial Plan, one Spatial Plan of the Unit of Local Administration and four Spatial Plans for Special Purpose Areas. The analysis and annotation of the texts, coupled with various methodological approaches, were applied on the subcorpus using language models, tools, and resources for the Serbian language developed within the Language Resources and Technologies Society – JeRTeh.

The prepared text of the plans, in the plain text format, was subjected to sentence segmenting using the tool Unitex, followed by word recognition in the text using the existing Morphological Electronic Dictionary for the Serbian Language. Further annotation of the subcorpus of spatial plans included the application of the named entity recognition system SrpNER, which automatically marks geopolitical terms, organization names, and demonyms. The first two classes of named entities were further subdivided into subclasses. For example, organization names were classified into commercial, political, religious, and general organizations, or organizations that were not otherwise classified.

The work further presents the INCEpTION tool, which allows the correction of automatically annotated named entities and their linking to the Wikidata knowledge base. In order to enable this linkage, the corresponding recognized named entities needed to have entries in Wikidata. This required additions to the knowledge base by creating missing entries and supplementing existing entries with missing data, i.e., properties. Besides posing various queries and listing named entities according to predefined criteria within the INCEpTION system, formation of SPARQL queries in the Wikidata knowledge base allowed different forms of visualization for the obtained results, i.e., in tabular form, as graphs, or as geographical maps.

Further research will focus on expanding the existing exemplary corpus of spatial planning and its subcorpus of spatial plans, as well as on improving and adapting the SrpNER named entity extraction system to the specific characteristics of planning documents.

Keywords: *digital corpus, spatial plans, corpus annotation, named entities, SrpNER, INCEpTION, Wikidata*

Acknowledgment: *This research was supported by the Science Fund of the Republic of Serbia, #7276, Text Embeddings - Serbian Language Applications - TESLA.*

References

[1] Bianchini, Carlo, Stefano Bargioni, and Camillo Carlo Pellizzari di San Girolamo. (2021). “Beyond VIAF: Wikidata As a Complementary Tool for Authority Control in Libraries”. *Information Technology and Libraries* 40 (2). https://doi.org/10.6017/ital.v40i2.12959

[2] Frontini, F., Brando, C., Byszuk, J., Galleron, I., Santos, D., Stanković, R. (2021). Named Entity Recognition for Distant Reading in ELTeC. In Constanza Navarretta; Maria Eskevich (ed.), *CLARIN Annual Conference 2020,* 36-41. http://hdl.handle.net/10400.26/39114.

[3] Ikonić Nešić, M., Stanković, R., & Rujević, B. (2022). Serbian ELTeC Sub-Collection in Wikidata. *Infotheca - Journal for Digital Humanities*, 21(2), 60-87. https://doi.org/10.18485/infotheca.2021.21.2.4

[4] Krstev, Cvetana, Ivan Obradović, Miloš Utvić, and Duško Vitas. (2014). A system for named entity recognition based on local grammars. *Journal of Logic and Computation*, 24 (2): 473–489.

[5] Milinković, M. (2022). Application of TXM Tools for Spatial Plan Corpus Analysis. *Infotheca - Journal For Digital Humanities, 22*(1), 32-51. https://doi.org/10.18485/infotheca.2022.22.1.2

[6] Milinković, M. (2022a). *The development of library and language resources for organizing and finding information on spatial planning*. PhD dissertation. University of Belgrade - Faculty of Philology. https://hdl.handle.net/21.15107/rcub_raumplan_683

[7] Nielsen, F.Å., Mietchen, D., Willighagen, E. (2017). Scholia, Scientometrics and Wikidata. In: Blomqvist, E., Hose, K., Paulheim, H., Ławrynowicz, A., Ciravegna, F., Hartig, O. (eds) *The Semantic Web: ESWC 2017 Satellite Events. ESWC 2017. Lecture Notes in Computer Science*. Vol 10577. Springer, Cham. https://doi.org/10.1007/978-3-319-70407-4_36

[8] Stanković, R. (2022). Distant Reading Training School 2020: Named Entity Recognition & Geo-Tagging for Literary Analysis. *Infotheca - Journal For Digital Humanities, 21*(2), 167_171. Retrieved from https://infoteka.bg.ac.rs/ojs/index.php/Infoteka/article/view/223

[9] van Veen, Theo. (2019). Wikidata: From “an” Identifier to ‘the’ Identifier. *Information Technology and Libraries* 38 (2):72-81. https://doi.org/10.6017/ital.v38i2.10886.

Ranka Stanković, Jovana Rađenović, Maja Ristić, Dragan Stankov
University of Belgrade, Faculty of Mining and Geology
E-mail: {ranka.stankovic|jovana.radjenovic|maja.ristic|dragan.stankov}@rgf.bg.ac.rs

Creation of a Training Dataset for Question-Answering Models in Serbian

The development and application of artificial intelligence in language technologies have advanced significantly in recent years, especially in the domain of the task of answering questions (Question Answering - QA). While existing resources for QA tasks have been developed for major world languages, the Serbian language has been relatively neglected in this area. This work represents an initiative to create an extensive and diverse set of data for training models for answering questions in the Serbian language, which will contribute to the improvement of language technologies for the Serbian language.

In addition to the numerous research on language models in the last few years, much work has also been done on the reference datasets needed to track modeling progress. A lot has been done when it comes to answering questions and understanding what is read, although, mostly, when it comes to big languages (Rogers et al. 2023). The paper provides an overview of the various formats and domains of available multilingual and monolingual resources, with special reference to the Serbian language (Cenić & Stojković 2023; Cvetanović & Tadić 2024). We will also consider the implications that follow from an excessive focus on the English language

As part of the TESLA (Text Embeddings - Serbian Language Applications) project, we are working on the preparation of a set of data: context, questions and answers, collected from different domains. The set will be made up of three smaller ones. To create the first set, a subset of the Stanford set SQuAD (Rajpurkar et al. 2018), where the answer is a segment of text, is translated and adapted, choosing topics such as: Nikola Tesla, climate change, construction, geology, etc. The subset will have around 7000 questions with accompanying answers. The second set that is being prepared will mainly be related to environmental protection, informatics and energy and will contain about 5000 questions with answers and given context excerpted from the textbook. The third set will contain automatically generated contexts based on the content of the Wikidata knowledge base.

The questions are carefully formulated to cover different types of queries: questions that require specific facts, questions with descriptive answers (which seek explanations or descriptions), and procedural questions, that is, questions that require a series of instructions or steps as a response. Data is collected in a variety of ways and verified through a manual annotation process to ensure accuracy and relevance of responses. The lack of manually annotated datasets in the Serbian language makes the contribution of this research particularly important.

The conclusion of the paper indicates the importance and potential of the application of this data set in various fields, including educational technologies, digital assistants, and information retrieval systems. The presented results contribute to the improvement of language technologies for the Serbian language, and we hope that they will encourage further research and development in this area.

Keywords: *artificial intelligence, natural language processing, language resources, annotated sets, information extraction, question answering*

Acknowledgment: *This research was supported by the Science Fund of the Republic of Serbia, #7276, Text Embeddings - Serbian Language Applications - TESLA.*

References

[1] Rogers, Anna, Matt Gardner, and Isabelle Augenstein. "QA Dataset Explosion: A Taxonomy of NLP Resources for Question Answering and Reading Comprehension." ACM Computing Surveys 55, no. 10 (2023): 1-45. https://arxiv.org/pdf/2107.12708

[2] Rajpurkar, Pranav, Robin Jia, and Percy Liang. "Know what you don't know: Unanswerable questions for SQuAD." arXiv preprint arXiv:1806.03822 (2018). https://arxiv.org/abs/1806.03822 , https://rajpurkar.github.io/SQuAD-explorer/

[3] Cenić, Aleksandar B., and Suzana Stojković. "A Serbian Question Answering Dataset Created by Using the Web Scraping Technique." In 2023 58th International Scientific Conference on Information, Communication and Energy Systems and Technologies (ICEST), pp. 147-150. IEEE, 2023.

[4] Cvetanović, Aleksa, Predrag Tadić, “Synthetic Dataset Creation and Fine-Tuning of Transformer Models for Question Answering in Serbian”, 2024, https://arxiv.org/html/2404.08617v1, https://paperswithcode.com/paper/synthetic-dataset-creation-and-fine-tuning-of

Jasmina Moskovljević Popović
Univerzitet u Beogradu - Filološki fakultet
E-mail: jasmina.moskovljevic@fil.bg.ac.rs

Challenges in Developmental Data Annotation

The paper reports on the difficulties and challenges encountered in planning and conceptualizing an annotation and marking strategies and procedures to be performed on the developmental corpus of Serbian language (RAKOPS). The RAKOPS is a representative developmental corpus of Serbian written language, created and compiled at the Department of General Linguistics and the Centre for Applied Linguistics of the Faculty of Philology, University of Belgrade over the last decade and a half. The entire content of RAKOPS, nearly 7000 texts of different genres authored by primary school pupils aged 8 to 15 years, has been digitalized and is available in two different formats – .html, and .txt, making it suitable for various types of analyses.

As annotating the developmental corpus present significant challenges when it comes both to manual and automated annotation, an overview of the main concepts relating to planned procedures is provided, along with some discussion of the practical aspects and difficulties encountered. For the purposes of illustration, the RAKOPS data from texts written by children of different ages have been used and analysed. The focus has been on the discussion of the optimal procedures which may be used on texts which often enough don’t follow established orthographic and segmentation conventions, and which quite frequently swarm with errors. All main types of linguistic annotation have been considered: (1) part-of-speech (POS) tagging, (2) lemmatization, (3) syntactic parsing, (4) semantic annotation, and (5) pragmatic annotation.

Keywords: *written language, developmental corpus, RAKOPS; linguistic annotation, error annotation*

Saša Marjanović
University of Belgrade – Faculty of Philology
E-mail: sasa.marjanovic@fil.bg.ac.rs

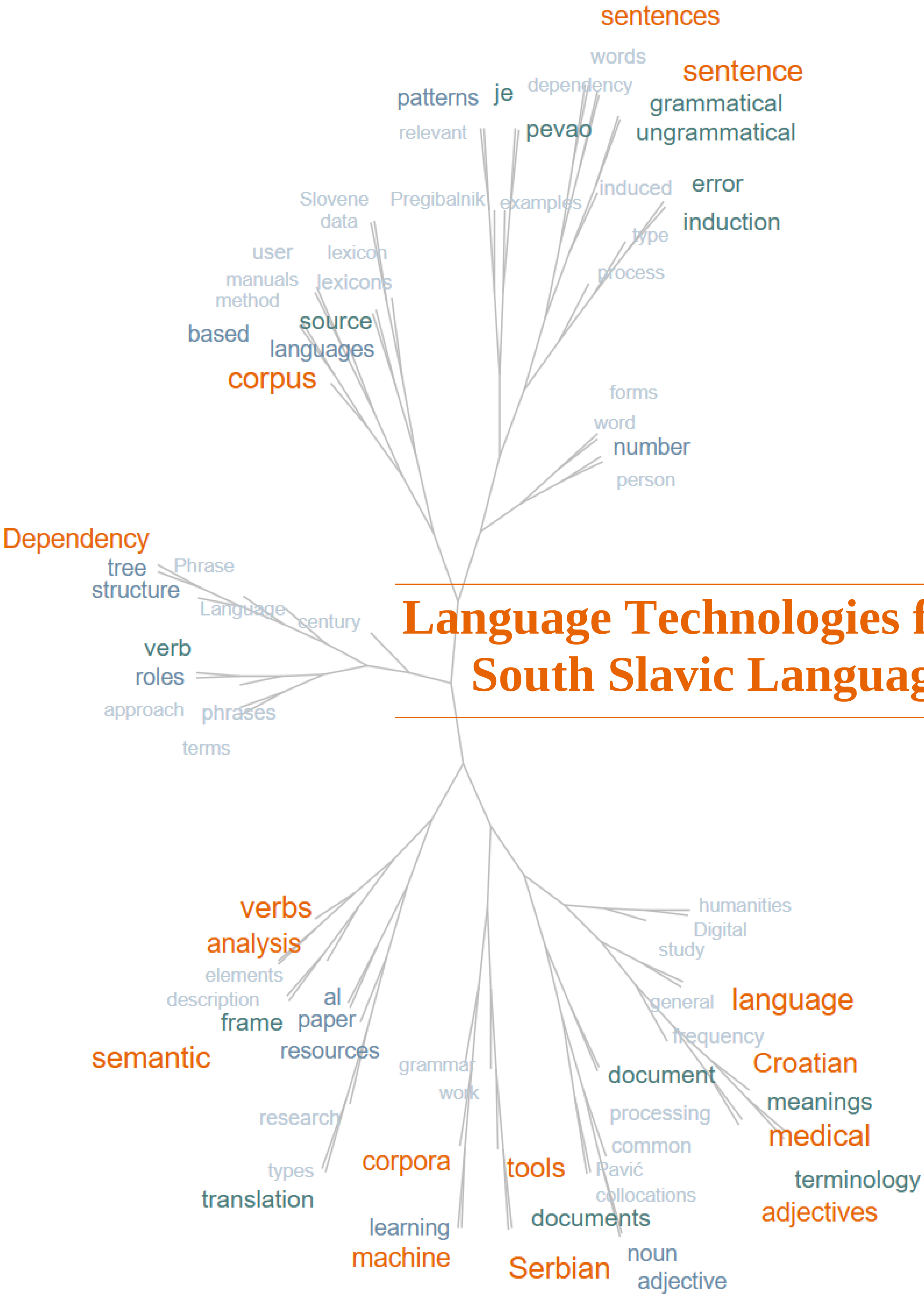
Dejan Stosic
University of Toulouse Jean Jaurès & CLLE (UMR 5263 CNRS)
E-mail: dejan.stosic@univ-tlse2.fr

The Application of The Electronic Conjugator *Serboverb* in the Study of Acquisition of Verb Inflection in the Serbian Language

The electronic conjugator *SerboVerb* (Stosic et al. 2018) is a linguistic resource for the Serbian language created at the University of Toulouse – Jean Jaurès (France) in collaboration with researchers from the Faculty of Philology at the University of Belgrade. The resource is available free of charge through both the web application (accessible at: https://serboverb.com) and mobile applications for *Android* and *iPhone*. It is primarily intended for users who are learning Serbian as a second language (including foreign, non-native, and hereditary speakers), but it can also be used by native Serbian speakers of all ages. The resource consists of three modules: conjugation, dictionary, and gamification modules. The first two modules are internally linked and form the core of the resource, while the third is a collection of externally stored tests for the additional application of the first two modules. The conjugation module is a searchable, exhaustive morphological lexicon that includes 34,000 lemmatized verbs of the Serbian language, along with their associated simple and compound inflectional forms. So far, over 20,000 verbs have been manually checked and verified. In the dictionary module, basic equivalents in other languages have been added to the verbs from the conjugation module. Currently, equivalents in ten languages are included for a set of 1,800 verbs that are mastered at higher levels according to the Common European Framework of Reference for Languages (CEFR). Additionally, equivalents in a total of 36 languages are provided for the hundred most frequently looked-up verbs in the conjugation module.

The functions of the entire resource are, therefore, multiple (cf. Marjanović et al. 2023). First, it can be useful in communication situations: on one hand, it helps users understand each individual form of a verb and its connection to the canonical form; on the other hand, it aids in producing any form within the verb paradigm, starting from any form known to the user. Additionally, it can be used in cognitive situations for learning and acquiring individual inflectional forms and verb inflections of the Serbian language in general. Given that the web application of this resource includes an administrative panel that collects and displays various usage data, administrators have the opportunity to analyze this data to draw conclusions that may be important for designing the teaching methods for verb inflection in Serbian as a second language. This presentation first demonstrates how the look-up frequency of certain verbs, cross-referenced with their frequencies from a general corpus of the Serbian language and verb lists from textbooks for Serbian as a foreign language, outlines the tendencies in the distribution of verbs according to CEFR levels. It then shows how the frequency of look-ups for verbs and their inflectional forms highlights the primary conjugation patterns for acquiring Serbian as a second language. Finally, it indirectly indicates the difficulties users may face in acquiring verb inflection in Serbian.

Keywords: *SerboVerb, verb, inflection, conjugator, second language acquisition*



Language Technologies for South Slavic Languages

Marija Đokić Petrović
Virtual Vehicle Research GmbH
Inffeldgasse 21a, 8010, Graz, Austria
E-mail: office@marijadjokicpetrovic.com

Mihailo St. Popović
Austrian Academy of Sciences, Institute for Medieval Research
Georg-Coch-Platz 2, 1010, Wien, Austria
E-mail: mihailo.popovic@oeaw.ac.at

Vladimir Polomac
University of Kragujevac, Faculty of Philology and Arts
Jovana Cvijića bb, 34000, Kragujevac, Serbia
E-mail: v.polomac@filum.kg.ac.rs

Utilizing Named Entity Recognition for the Analysis of Serbian Archival Documents

Digital humanities represent a significant advance in the research and preservation of cultural heritage, enabling researchers to use modern technologies to analyze and interpret historical documents. Our paper analyzes two Serbian archival documents written in diplomatic minuscule Cyrillic. The first document from 1778 is unpublished and preserved in the archives of the Greek Church of St. George in Vienna, the oldest Orthodox church in today's Austria. The second document from 1500 is the testament of Miloš Belmužević, a Serbian lord serving the Hungarian kings. The documents were initially subjected to manual reading and examination to extract information about persons, locations, and demographics. Simultaneously, a computational approach utilizing the Named Entity Recognition (NER) was employed to accomplish the same task. The dataset, which comprised the Charter of King Stefan Uroš II Milutin to the Monastery of St. Stefan in Banjska from the beginning of the 14th century, the Dečani chrysobulls from the 14th century, and the collection of 13th-century charters and letters from the Dubrovnik archive, was used to train the NER model. At the end, a comparative analysis of the traditional approach and the results of computational processing was performed, facilitating the assessment of the efficiency of both methods. This analysis additionally confirmed the value of Digital humanities in the preservation and study of Serbian historical documents.

Keywords: *Named Entity Recognition; Digital humanities; Diplomatic minuscule Cyrillic; Cultural heritage; Serbian archival documents; Historical preservation*

References

[1] Todorović, B. Š., Krstev, C., Stanković, R., & Nešić, M. I. (2021, September). *Serbian NER&Beyond: The archaic and the modern intertwinned*. In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021) (pp. 1252-1260).

[2] Cvejić, A. (2022). *Analiza različitih modela za Prepoznavanje Imenovanih Entiteta na srpskom jeziku*. Zbornik radova Fakulteta tehničkih nauka u Novom Sadu, 37(02), 316-319.

[3] Jehangir, B., Radhakrishnan, S., & Agarwal, R. (2023). *A survey on Named Entity Recognition—datasets, tools, and methodologies*. Natural Language Processing Journal, 3, 100017.

Nikola Janković
University of Belgrade - Faculty of Philology,
E-mail: nikolajankovickv@gmail.com

Methodology for Creating a Multilingual Parallel Corpus Based on Online Digital User Manuals: The Hilti User Manuals Corpus

This paper presents a methodology for creating an extensive multilingual parallel corpus of user manuals, specifically based on digital user manuals of the company Hilti. The goal of this paper is to enhance research resources in the fields of machine translation, translation theory, as well as translation teaching and other types of translation research. The starting point of the methodology was the uniform HTML structure of the user manuals for all languages on the official Hilti online database of user manuals. Further steps involved identifying HTML elements which contain the relevant segments of texts, their consistent numeration for each language, identifying and cleaning incorrectly parallelized texts, and storing the processed texts in TMX format, so that English is the source language in each file. On average, for each user manual there are around 20 target languages, and the entire corpus contains around 7 million words. The advantages of this corpus include a high level of precision when it comes to segment matching across different languages, a substantial word count, as well as the significant number of languages contained in the corpus. The paper will describe in detail the steps of scraping, processing, and structuring the data, the technical challenges faced in the process and their solutions, as well as the basic statistical data related to the languages in the corpus. We believe that this type of corpus can be used for improvement of translation resources, translation theory, translation teaching, as well as for different types of analyses of specialized language.

Keywords: *parallel corpora, machine translation, language tools, specialized language, TMX*

Anđelka Zečević, Anastazia Žunić, Kristina Milojević
Mathematical Institute, Serbian Academy of Sciences and Arts
E-mail: {andjelkaz|anastazia.zunic|kristinam}@mi.sanu.ac.rs

Old Texts, New Technologies: Digitization of Documents in Serbian

Numerous tools and platforms are emerging in the digital humanities, intending to advance and enrich work with digitized documents. At the core of these tools are complex machine learning algorithms for image and text processing, trained on appropriately prepared document repositories. Their most common functionalities include image quality enhancement, document layout analysis, optical character recognition, and text post-correction. Newer tools also offer support for standardizing orthography, convenient and more comprehensive document retrieval, toponyms extraction, topic recognition, and document summary generation.

This work will present experiences collected using modern open-source tools for digitizing documents in Serbian. The tools in question are *Calamari-OCR*, *docTr*, *LayoutParser*, *Kraken*, *Tesseract*, *OCR4All*, and others, designed to address individual steps in the digitization process or the complete *end-to-end* pipeline. The resources used for testing these tools will be periodicals published during the 19th century provided by the National Library of Serbia, characterized by a great diversity of graphic elements, non-standard formats, physical degradation, and lower-quality scans. Along with the observed advantages and limitations of these tools, the work will also discuss ways to further expand and adapt them to the Serbian language. Special attention will be given to the role of language technologies and scenarios where their use is essential.

Keywords: *document digitization, periodicals, image processing, natural language processing*

Nikitas N. Karanikolas, Professor
Department.of Informatics & Computer Engineering, University of West Attica, Greece
E-mail: nnk@uniwa.gr

Noun Phrase and Prepositional Phrase Chunking for the Greek Language with Spacy

In the need for Natural Language Understanding (NLU) various tasks should be done. As a short abstraction of this process, the text sentences should (a) be syntactically analysed and (b) some of the phrases constituting the syntactic (Phrase structure) tree should be assigned to the semantic/thematic slots/roles that correspond to the frame (case grammar) of the sentence's verb (event). The most promising phrases to fill the semantic/thematic slots/roles of a verb (event) are noun phrases (np) and prepositional phrases (pp). Consequently, the parser (the Syntactic analyzer | the Phrase Structure analyzer), either a deep or a shallow parser, should be able to isolate np and pp phrases.

An alternative approach for the same step (a) is to use a Dependency analyzer and extract the Dependency structure/tree for each sentence. The Dependency structure/tree of a sentence provides the dependency (grammatical) relations between the words forming a sentence. A dependency tree depicts the head words (up) and the dependent words (down). The head word at the top is the Root of the Dependency tree. The arcs linking the vertices are labeled with the dependency (grammatical) relations. The advantage of Dependency trees is that the arguments to the verb are directly linked to it in the Dependency tree.

The resources needed for NLU are various and most of them are large, populous/crowded, demanding. We only mention: Lexicons able to recognize all the inflected forms of a word and return the head (lemma) and the features (gender, number, and case for nouns, voice, tense, mood, number, and person for verbs, etc); Collections of Frames for verbs (events) containing the relevant to the verb Thematic roles, Selectional restrictions for Thematic roles, etc.

There are tools that provide abilities for Syntactic (Phrase structure) or Dependency analysis that relieve us of the need to have our own Lexicons. However, it is not clear if these tools are able to provide all the promising phrase structures (np, pp, etc) needed to fill the Thematic roles and also check the Selectional restrictions imposed by Frames. We are evaluating such a tool for the Greek Language.

Keywords: *Greek language, extracting phrase structures, dependency parsing*

References

[1] Daniel Jurafsky & James H. Martin, Speech and Language Processing, 2019. Chapter 15 Dependency Parsing. https://web.stanford.edu/~jurafsky/slp3/old_oct19/15.pdf

[2] Fei Xia & Martha Palmer, Converting Dependency Structures to Phrase Structures. Human Language Technology - The Baltic Perspectiv, 2001. https://aclanthology.org/H01-1014.pdf

[3] Nikitas Karanikolas et al, Large language models versus natural language understanding and generation. ACM Digital Library, pub. 14 Feb. 2024. https://doi.org/10.1145/3635059.3635104

Jaka Čibej

Faculty of Arts, University of Ljubljana
Centre for Language Resources and Technologies, University of Ljubljana
E-mail: jaka.cibej@ff.uni-lj.si

First Steps Towards an Online Service for Automatic Morphological Inflection of Serbian and Croatian

Open-source machine-readable morphological lexicons are useful for morphosyntactic tagging of corpora and represent a crucial step toward compiling modern digital dictionary databases (see e.g. Kosem et al. 2021). Currently, the most well developed among the lexicons for South Slavic languages is the *Sloleks Morphological Lexicon of Slovene* (Čibej et al. 2022). Version 2.0 with approximately 100,000 entries was updated to version 3.0, adding approximately 265,000 new entries, their inflected forms, accentuated forms, and IPA/SAMPA pronunciations. All were automatically generated using *Pregibalnik* ("Inflector"), a custom-developed open-source tool (also available as an API service[1]) for Slovene lexicon expansion, which takes a lemma and its morphosyntactic features according to the MULTEXT-East Morphosyntactic Specifications[2] (e.g. *liofilizacija*, 'lyophilization'; noun, common, feminine) as input and generates (among other things) complete paradigms of forms inflected by case, number, tense, etc. (e.g. *liofilizacija*, *liofilizacije*, *liofilizaciji*, ...) using a combination of machine-learning and linguistically informed rule-based methods, including machine-readable morphological patterns (e.g. "[*liofilizacij*]-a, [*liofilizacij*]-e, [*liofilizacij*]-i, ...") which were automatically extracted and validated (Arhar Holdt & Čibej 2018; Arhar Holdt 2021) before being used in machine-learning predictions.

Two open-source lexicons similar to Sloleks have been published for Serbian and Croatian – srLex 1.3 (Ljubešić 2019a) and hrLex 1.3 (Ljubešić 2019b), compiled from srWaC and hrWaC corpora, respectively. However, while machine-learning methods for lexicon expansion have already been used to predict paradigms for Croatian and Serbian (e.g. Ljubešić et al. 2016; also Šnajder 2013), the paradigms were only available in the Apertium format,[3] which is not compatible with the *Pregibalnik* infrastructure. The patterns need to be converted and cross-

[1] https://orodja.cjvt.si/pregibalnik/redoc
https://orodja.cjvt.si/pregibalnik/docs
https://orodja.cjvt.si/pregibalnik/form-generator/docs
https://orodja.cjvt.si/pregibalnik/form-generator/redoc

[2] MULTEXT-East Morphosyntactic Specifications for Slovene: https://nl.ijs.si/ME/V6/msd/html/msd-sl.html

[3] https://sourceforge.net/p/apertium/svn/HEAD/tree/languages/apertium-hbs/apertium-hbs.hbs.metadix

checked with srLex and hrLex in order to be successfully implemented in an easily accessible API service. Because Croatian and Serbian are structurally similar to Slovene and share a similar infrastructural framework, the same method applied to Slovene data can be used (with some minor adjustments) to implement the patterns into *Pregibalnik*. This will not only be the first step toward extending the functionalities of *Pregibalnik* to cover Serbian and Croatian and help automatically expand the lexicons with new entries, but will also identify potential inconsistencies within the current versions of the lexicons.

Acknowledgment: *The author acknowledges the financial support from the Slovenian Research and Innovation Agency (research core funding No. P6-0411 – Language Resources and Technologies for Slovene).*

Keywords: *lexicon, morphology, inflection, expansion, Croatian, Serbian*

References

[1] Arhar Holdt, Špela & Jaka Čibej, 2018, Oblikoslovni vzorci v leksikonu Sloleks: izhodiščni nabor za samostalnike. In: *Slovenščina 2.0: Empirične, Aplikativne in Interdisciplinarne Raziskave, 6(2)*, pp. 33-66. https://doi.org/10.4312/slo2.0.2018.2.33-66

[2] Arhar Holdt, Špela, 2021. Oblikoslovni vzorci za strojno procesiranje slovenščine. In: Arhar Holdt, Špela (ed.): *Nova slovnica sodobne standardne slovenščine: viri in metode*. Založba Univerze v Ljubljani. https://doi.org/10.4312/9789610605478

[3] Čibej, Jaka, Kaja Gantar, Kaja Dobrovoljc, Simon Krek, Peter Holozan, Tomaž Erjavec, Miro Romih, Špela Arhar Holdt, Luka Krsnik, Marko Robnik-Šikonja, 2022, *Morphological lexicon Sloleks 3.0*, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1745

[4] Kosem, Iztok, Simon Krek, Polona Gantar, 2021, Semantic data should no longer exist in isolation: the Digital Dictionary Database of Slovenian In: *EURALEX XIX, Congress of the European Association for Lexicography, Lexicography for inclusion*, pp. 81–83.

[5] Ljubešić, Nikola, 2019a, *Inflectional lexicon srLex 1.3*, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1233.

[6] Ljubešić, Nikola, 2019b, *Inflectional lexicon hrLex 1.3*, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1232.

[7] Ljubešić, Nikola, Filip Klubička, Željko Agić, Ivo-Pavao Jazbec, 2016, New Inflectional Lexicons and Training Corpora for Improved Morphosyntactic Annotation of Croatian and Serbian. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 4264–4270, Portorož, Slovenia. European Language Resources Association (ELRA).

[8] Šnajder, Jan, 2013, Models for predicting the inflectional paradigm of Croatian words. In: *Slovenščina 2.0, 1 (2)*, pp. 1–34. http://www.trojina.org/slovenscina2.0/arhiv/2013/2/Slo2.0_2013_2_02.pdf.

Martina Pavić
Institute for the Croatian Language
E-mail: mpavic@ihjj.hr

Frequency of Adjective Meanings in Croatian Medical Terminology – A Corpus-Based Analysis of Collocations

Adjectives are an important component of Croatian medical terminology, contributing to its high level of structuring by narrowing and modifying the meaning of the words they are attached to. They can be crucial in the formation of conceptual systems (Pitkänen-Heikkilä 2015, see Pavić 2022: 153).

The frequency of adjective meanings in medicine reflects the way health conditions, diagnoses, and treatments are described. Researching this frequency is essential for the standardization and precision of medical terminology, which is extremely important for clarity and safety in medical communication and practice. In the medical field, this part of speech is involved in the classification and specification of diseases, diagnoses, symptoms, indications, and other conceptual categories. Considering that the semantic realization of adjectives in medical terminology should be observed in usage, this paper will apply a corpus-based research model, using a scientific corpus of contemporary medical journals, comprising 5 318 395 tokens, compiled in the *Sketch Engine* tool (Pavić 2022). In this study, we will apply a classification of 14 meanings of adjectives in Croatian medical terminology – function, cause, time/duration, physical feature, composition, state, property, location, outcome, application, source, age, intensity, relationship – proposed in Pavić (2022). In the semantic analysis, we will start with the highest categories of terms, using a top-down approach. The categories of terms from Taylor (2017) such as <disease>, <symptom>, <indication>, <diagnosis>, and <procedure> were adopted, and after extensive corpus analysis, the following were added: <medical staff>, <patient>, <side effect>, <body part>, and <medical equipment>. Taylor's category of <drug> will be included in the <therapy> category, as therapy is most often lexicalized by the noun *drug* and its types (Pavić 2022).

For the first time, we will investigate the frequency of adjective meanings in Croatian medical terminology, extracted using a semi-automated method in the *Keywords* option (with a frequency greater than 200). We will start from the list of the most common adjective-noun collocations (Pavić 2022), as the assumption is that an adjective expresses different features depending on the noun it modifies (Grčić Simeunović 2021). In the collocations, we will determine which meaning or meanings the adjectives express (for example, the adjective *hormonal* expresses four meanings in different adjective-noun combinations). In addition to adjectives with inherently medical meanings (e.g., *blood*, *endoscopic*, *gastrointestinal*), we will also investigate the frequency of meanings of adjectives from general language that form medical terms in multi-word expressions (e.g., *right*, d*eep*, *central*).

The goal of this study is to use a corpus-based method to determine the most common meanings of adjectives in Croatian medical terminology, examine which adjectives in collocations change the category of terms, and whether their meaning changes accordingly. By including adjectives from general language in the semantic analysis, the study will also examine the relationship between general and specialized language, as well as the existence of any semantic shift.

Keywords: *adjectives, semantic analysis, Croatian medical terminology, scientific corpus, corpus-based method, collocations*

References

[1] Grčić, Simeunović, L. (2021) *Terminološki opis u službi stručnoga prevođenja. Dinamično modeliranje specijaliziranoga znanja*. Zadar – Zagreb: Sveučilište u Zadru – Institut za hrvatski jezik i jezikoslovlje.

[2] Pavić, M. (2022) *Uloga pridjeva u hrvatskome medicinskom nazivlju*. Doktorska disertacija. Sveučilište u Zagrebu, Filozofski fakultet, Zagreb.

[3] Pitkänen-Heikkilä, K. (2015) Adjective as terms. *Terminology,* 21 (1), 76–101.

[4] Taylor, B. R. (2017) *The Amazing Language of Medicine: Understanding Medical Terms and Their Backstories*. New York City: Springer International Publishing.

Ana Ostroški Anić, Ivana Brač

Institute for the Croatian Language, Zagreb, Croatia

E-mail: aostrosk@ihjj.hr

A Frame-Semantic Description of Cognition Verbs in Croatian

Semantic description of verbs in many lexical resources predominantly includes the organization of verbs into semantic classes, commonly following the seminal work by Levin (1993), without a deeper semantic analysis.

The SEMTACTIC research project (https://semtactic.jezik.hr/) aims to go beyond its central focus of determining the semantic classes of the 500 most frequent verbs in the Croatian language by exploring their prototype syntactic patterns and semantic roles. The relation between the semantics and syntax of Croatian verbs will be explored both within a semantic class and across classes by means of different descriptions, one of which is the frame-semantic framework used to define verbs following the principles of Frame Semantics (Fillmore 1985; Fillmore, Johnson and Petruck, 2003; Ruppenhofer et al., 2016).

This paper presents the frame-semantic description of the verbs of cognition in the Verbion database, which is being developed within the project. The argument structure of verbs like *misliti* 'think', *smisliti* 'come up with', *promisliti* 'think through', *razmišljati* 'think', *predložiti* 'suggest', etc. is explained, along with their semantic description within the appropriate semantic frames of FrameNet, e.g. `Opinion`, `Judgement` or `Coming_up_with`. Relevant corpus examples of sentences in which these verbs are used as target lexical units are annotated for frame elements in order to compare the verbs according to their valence frames. E.g., in the sentence [Ovu inovativnu uslugu]$_{\text{Idea}}$ SMISLILA je [tvrtka iz Arizone koja proizvodi prirodne preparate za njegu tijela]$_{\text{Cognizer}}$ ('A company form Arizona that produces natural body care products came up with this innovative service.'), the core elements of the frame `Coming_up_with` are annotated.

Defining cognition verbs following the methodology used in Framenet (Ruppenhofer et al., 2016) adds additional level of syntactic and semantic description in the Verbion database, as well as provides data for the creation of future frame-based lexicon of Croatian.

Keywords: *verbs, cognition verbs, FrameNet, frame semantics*

References

[1] Fillmore, Charles J. 1985. Frames and the semantics of understanding. *Quaderni di semantica: Rivista internazionale di semantica teorica e applicata* 6, 222–254.

[2] Fillmore, Charles J.; Johnson, Christopher R.; Petruck, Miriam R. L. 2003. Background to Framenet. *International journal of lexicography* 16/3, 235–250.

[3] Levin, Beth. 1993. *English Verb Classes and Alternations. A Preliminary Investigation*. The University of Chicago Press.

[4] Ruppenhofer, J. et al. 2016. *FrameNet II: Extended Theory and Practice*. https://framenet2.icsi.berkeley.edu/docs/r1.7/book.pdf

Marija Pantić

Society for Language Resources and Technologies JeRTeh
E-mail: p.mmmmmm@gmail.com

Artificial Corpora for Machine Training of Grammar Tools

Machine learning datasets for grammar checkers require large corpora of grammatical and ungrammatical sentences. Collecting naturally occurring ungrammatical sentences and classifying them by number and type of errors can be a complicated, time-consuming, and demanding task considering the necessary human and linguistic resources. Artificial error corpora can be a relatively simple and significantly cheaper, and often far more reliable, alternative, and they can be produced by an error induction process comprising minimal transformations in corpora of grammatical sentences with appropriate morphosyntactic tags. Corpora of grammatical sentences can be general or specifically built for the purpose of error induction. General corpora do not require additional preparation but must contain enough examples of relevant patterns, and the resulting corpus of induced ungrammatical sentences will be smaller than the source corpus of grammatical sentences. In contrast, if we select only sentences that contain patterns relevant to the induction of a chosen type of error, we ensure that each grammatical sentence from the source corpus will be transformed into an ungrammatical sentence in the error-induced corpus in accordance with the error-induction rules. When selecting the type of error for induction and writing induction rules, we should consider the following:

1. Error induction may not always produce an ungrammatical sentence.
   - *a)* *Pričam.* → *Priča.*
   - *b)* *Pričam.* → *Pričah.*
   - *c)* *Ja pričam.* → _*Ja priča._
   - *d)* *Pričam koliko hoću.* → *?Priča koliko hoću.*

In examples (1a) and (1b), replacing the first-person singular with any other person or replacing present tense with another simple personal verb form will not result in an ungrammatical sentence because the sentence consists only of a verb. In example (1c), a disagreement in person between the subject and the verb has been successfully induced. In example (1d), what has been induced is a grammatically possible but from the viewpoint of usage questionable sentence.

2. An ungrammatical sentence can often be corrected in several ways by an equally minimal number of transformations.
   - *a)* *On je pevao.* → On pevao. → *On je pevao; On bi pevao; On beše pevao: On peva; On pevaše.*
   - *b)* *On je pevao.* → On je pevala. → *Ona je pevala; On je pevao.*

In examples (2a) and (2b), the correct sentence does not offer the only or necessarily the best solution for correcting the error in the incorrect sentence.

This presentation will review some of the criteria for evaluating the quality of induced corpora, excessive and insufficient specificity of induction rules for different types of grammatical errors in Serbian, and some of the most common errors made by native and non-native speakers of Serbian that could have better coverage in grammar checkers for Serbian.

Keywords: *grammar tools, artificial corpora, machine learning*

Grammar and Lexicon of South Slavic Languages in the Context of NLP

Marina Bagi
Institute for Serbian Language SANU
E-mail: marina.bagi@isjsanu.ac.rs

Syntactic and Semantic Analysis of the Verbs *Oštetiti* and *Uništiti* from the Perspective of Frame Semantics Theory

This paper analyzes the verbs *oštetiti* ('to damage') and *uništiti* ('to destroy') from the perspective of the frame semantics theory. Starting from the assumption that syntactic phenomena correlate with semantic ones (and vice versa), the theoretical framework for this study is based on the theory of frame semantics, a cognitive-semantic approach developed by Charles Fillmore, which explicitly links the meaning of the word with the syntactic contexts in which it appears. The core thesis of this theory is that the meanings of sub-lexemes should be viewed in relation to semantic frames, which represent schematic depictions of conceptual structures and belief patterns underlying word meanings. This theory is fundamental to FrameNet, an electronic lexicographic database based on a corpus, within which the frames invoked by various lexemes are presented. One of the central principles of this approach is that certain types of words (verbs, nouns, adjectives, adverbs) evoke frames, which are conceptual systems connected in such a way that understanding one concept requires understanding the entire structure of which it is a part.

This study examines more than five hundred examples of the use of the verbal lexemes *oštetiti* ('to damage') and *uništiti* ('to destroy'), excerpted from the electronic Corpus of Contemporary Serbian (available on the platform [https://noske.jerteh.rs/]), with the aim of describing and examining their syntactic and semantic behaviour, as well as illustrating their usage. Consequently, the frames evoked by these verbs are described in detail, along with the elements of the frames, and conclusions are drawn regarding the syntactic-semantic interface. Furthermore, since these verbs share similar semantics, the aim is to investigate both the similarities and differences between them using tools available on the Sketch Engine and No Sketch Engine platforms. The analysis focuses on their most frequent collocates, the frequency of specific forms, word sketches, synonyms, and related aspects. The conducted research is part of the future, broader, ongoing study within a doctoral dissertation which deals with the application of the frame semantics theory in Serbian lexicography (on the example of the verbs associated with damage and destruction), as well as part of the project of creating the Serbian FrameNet. Another objective of the study is to outline the challenges encountered during this phase of the project.

Keywords: *frame semantics theory, syntax, semantics, verb oštetiti, verb uništiti, natural language processing, Serbian language*

Prof. dr Nataša Kiš
Univerzitet u Novom Sadu - Filozofski fakultet, Odsek za srpski jezik i lingvistiku
E-mail: natasab@ff.uns.ac.rs

Syntax-Semantic Annotation of Serbian Language Electronic Corpora

This paper will highlight the possibilities of using an electronic corpus as a primary source of material for various research in the fields of syntax and semantics of contemporary Serbian language. The two main tasks of the paper are to present the research conducted so far related to the process of complementation of adjectives and adjectival nouns, specifically their governing potential at the syntagmatic level, based on relevant material primarily sourced from the electronic *Corpus of Contemporary Serbian Language*, SrpKor2013 (www.korpus.matf.bg.ac.rs). In the analysis of the adjective complementation process alone, over 9000 sentences were excerpted from the corpus. The second task is to indicate the need for expanding corpus search capabilities from the perspective of sentence-level analysis. Searches of this type would include data on the valency properties of sentence predicates and all obligatory and optional arguments in a given situation, with an important segment of the analysis being the identification of semantic word classes and their syntax-semantic characteristics, which would allow for searches of various sentence models.

For example, the relational adjective *веран* (*faithful*) governs an object complement in the dative without a preposition marked by animate+/- (он је *веран* ***пријатељу*** / ***том идеологији*** – he is *faithful* ***to a friend / to that ideology***), while the noun *верност* (*faithfulness*) takes the same type of object complement realized in different forms (показао је *верност* ***пријатељу*** / ***према пријатељу*** – he showed *faithfulness* ***to a friend / towards a friend***), but can also be used without a complement in a certain context (они не могу да говоре ***о верности*** – they cannot talk ***about faithfulness***).

Special attention should be given to the polysemy of lexemes that also affects the governing potential of words. The adjective *веран* (*faithful*) and the noun *верност* (*faithfulness*), if the bearer of the characteristic has an animate- mark, can mean "being equal to something", implying that the complement they govern will belong to a different semantic type, i.e., it will denote the other bearer of the characteristic (текст је *веран* ***оригиналу***; издавач се држао *верности* ***према тексту*** – the text is *faithful* ***to the original***; the publisher adhered to *faithfulness* ***to the text***).

The corpus search would be more precise if various syntax-semantic data such as animacy, directionality, reciprocity, ablativeness, comprehensiveness, partitivity, etc., could be included as parameters. In sentence structuring, the issue of which linguistic means convey data about the agent/pseudo-agent is also relevant. In contemporary Serbian, for example, different structural types are distinguished based on these parameters (***ја*** *бих јела јабуке* / *једу* ***ми*** *се јабуке* – ***I*** *would eat apples* / ***I*** *feel like eating apples*, etc.).

From the perspective of a speaker of Serbian as a native or non-native/foreign language, it would be useful if searches of the electronic corpus could start from semantic word classes, whose formalization or structural models in which they are realized would be demonstrated by examples excerpted from the electronic corpus (e.g. in the class of words with the meaning of optative modality, there are the adjective *жељан* (*desirous*), the noun *жеља* (*desire*), and the verb *желети* (*to desire*), which can be realized in the following sentence models: Ја сам *жељна* ***одмора*** (I am *desirous* ***of rest***) – Моја *жеља* ***за одмором*** је велика / Ја имам *жељу* ***за одмором*** (My *desire* ***for rest*** is great / I have *a desire* ***for rest***)– Ја *желим* ***да се одморим*** (I *want* ***to rest.***).

These search parameters belong to the domain of syntax-semantic annotation of electronic corpora, making it crucial to highlight the appropriate linguistic analyses that would enable the extraction and inclusion of necessary data into the corpora themselves.

Keywords: *syntax, semantics, annotation of electronic corpora, governing ability, Serbian language*

Tanya Neycheva
Paisii Hilendarski University of Plovdiv
E-mail: neyche1va@uni-plovdiv.bg

Predicative Instrumental in Contemporary Serbian, Russian and Polish Languages (Based on Data Extracted from a Multilingual Online Dictionary)

The existing parallel language corpora enable comparative studies on material from two (rarely more) languages. Unfortunately, multilingual corpora are still not particularly rich and rely mostly on translated literary texts. In an attempt to find another reliable source of comparative research material, the author turns to multilingual online dictionaries. This type of platform relies on the so-called translation memory – a constantly updated database of translated texts, which allows the word (or whole word combination) searched for in the dictionary to be checked in numerous contexts in the original and in the translation language, and even in parallel in several languages. In this sense, the main feature that distinguishes a multilingual dictionary from a typical linguistic corpus is the lack of grammatical annotation (which, by the way, some corpora also lack).

The research presented here is dedicated to the Slavic predicative instrumental and was carried out on parallel examples from the Serbian, Russian and Polish languages extracted from the multilingual online dictionary Glosbe, and in the course of the work a simple algorithm was tested for manually discovering the necessary grammatical information in an unannotated corpus. The main steps include: selecting the most adequate keywords for the search, extracting the examples, removing the irrelevant and annotating the relevant results, analyzing the data.

As a result of the research, the peculiarities of the use of the predicative instrumental were described, as well as the way in which it competes with the concordant predicative cases (nominative, dative and partly accusative) in the contemporary Serbian, Russian and Polish languages. Several main subtypes of the compound predicate were examined - containing personal, impersonal and impersonal copula; transitive and intransitive (reflexive) personal semi-copula and impersonal semi-copula. Both the differences and some hitherto undescribed similarities between the three languages were highlighted.

The research also confirmed the assumption that multilingual online dictionaries, and more specifically Glosbe, can be used as a reliable source of data when conducting comparative linguistic research.

Keywords: *predicative instrumental, translation memory, unannotated corpus*

Svetlozara Leseva, Ivelina Stoyanova
Department of Computational Linguistics, Institute for Bulgarian Language, Bulgarian Academy of Sciences
E-mail: {zarka| iva}@dcl.bas.bg

Towards Classifying Activity Predicates Involving Change

The main objective of this paper is to offer a description framework of the semantic properties of activity (dynamic) predicates involving change, with the main focus on the property which undergoes the change. For our purposes, we understand *activity* in the widest possible sense, including different types of dynamic predicates.

The description of verbs of change is based on several key semantic features – the lexical meaning of the verb and the semantic class it belongs to, the type of change (causal or inchoative), the property of the change (quantised or non-quantised change, scalable change or momentous change of category or state), the frame elements of the relevant semantic frame in FrameNet (Baker and Fellbaum 2009; Ruppenhofer et al. 2016). The study relies on previous well-established semantic classifications of verbs, and focuses on the verbs in the Princeton WordNet (Miller 1995) and the Bulgarian WordNet (Koeva 2021).

Based on the outlined semantic features, a shallow classification within the semantic class of verbs of change will also be offered, aiming to cover the diversity within the class and to propose an approach to model their syntactic behaviour. The classification is based on FrameNet frames which group together verbs with similar conceptual properties and syntactic behaviour. The semantic properties determine to a large degree the syntactic realisation of the predicates in terms of the semantic frame they evoke as well as the configuration of frame elements and the way they are realised.

Our observations are based on automatically extracted and manually selected illustrative examples for Bulgarian and English. The English examples and annotations are borrowed from the FrameNet corpus, while the Bulgarian ones are manually annotated. For both languages the data are supplemented with examples from other corpora, where needed.

Annotated example:

FRAME: Cause_expansion; Default: [causal gradual change]
[Universitetat]$_{\text{AGENT}}$ postoyanno **razshiryava** [uchastieto si v kulturniya zhivot]$_{\text{ITEM}}$.
[The University]$_{\text{AGENT}}$ constantly **expands** [its participation in the cultural life]$_{\text{ITEM}}$.

FRAME: Cause_temperature_change; Default: [inchoative gradual change]
[Plamnaloto i litse]$_{\text{ITEM}}$ **se ohladi** [ot ledenata voda]$_{\text{CAUSE}}$.
[Her flushed face]$_{\text{ITEM}}$ **cooled** [by the chill water]$_{\text{CAUSE}}$.

We explore the universal aspects of conceptual knowledge that enable the transfer of semantic and syntactic information across resources and languages. The configuration of frame elements describing the behaviour of verbs evoking a particular frame are language-independent, as well as the semantic restrictions determining their selection. The constellations of frame elements that get syntactically expressed in combination with each other (i.e. the so-called valence patterns in FrameNet) are largely valid across languages as attested for Bulgarian and English; we also find substantial correspondence in the syntactic categories and grammatical functions by which frame elements are expressed for these languages.

We also analyse the language-specific properties of the semantic and syntactic description and the differences between English to Bulgarian based on empirical material from corpora. We view cases where the two languages give different preference to the overt expression of particular frame elements, or realise them in different syntactic positions. The observations can be informative for other Slavic languages exhibiting similar grammatical specificity.

Keywords: *verb semantics, frame semantics, aspectual classes, activity predicates, verbs of change*

References

[1] Colin F. Baker and Christiane Fellbaum. 2009. WordNet and FrameNet as Complementary Resources for Annotation. In Proceedings of the Third Linguistic Annotation Workshop (ACL-IJCNLP ’09), Association for Computational Linguistics, Stroudsburg, PA, USA, pages 125–129.

[2] Svetla Koeva. 2021. The Bulgarian WordNet: Structure and specific features. Papers of Bulgarian Academy of Sciences, 8(1):47–70.

[3] George A. Miller. 1995. WordNet: A lexical database for English. Commun. ACM, 38(11):39–41.

[4] Josef Ruppenhofer, Michael Ellsworth, Miriam R. L. Petruck, Christopher. R. Johnson, Collin. F. Baker, and Jan Scheffczyk. 2016. FrameNet II: extended theory and practice. International Computer Science Institute, Berkeley, California.

Maja Matijević
Institut za hrvatski jezik
E-mail: mmatijevic@ihjj.hr

From Dictionaries and Corpora to Hyponymy and Meronymy

Hyponymy and meronymy are hierarchical lexical-semantic relations that organize lexical hierarchies. Hyponymy is a lexical-semantic relation in which the unit on a higher hierarchical level (hypernym) denotes a type, while the units on a lower hierarchical level (hyponyms) represent specific instances of that type. For example, cvijet (‘flower’) is a hypernym, while ruža (‘rose’), tulipan (‘tulip’), and zumbul (‘hyacinth’) are its hyponyms. Meronymy, on the other hand, is a lexical-semantic relation in which the unit on a higher level (holonym) denotes a whole, while the units on a lower level (meronyms) represent parts of that whole. For instance, tijelo (‘body’) is a holonym, and ruka (‘hand’) and stopalo (‘foot’) are its meronyms. When defining hyponyms and meronyms, established patterns are commonly used, and lexicographic definitions, as clearly structured and harmonized components of dictionary entries, serve as a good source for identifying such patterns.

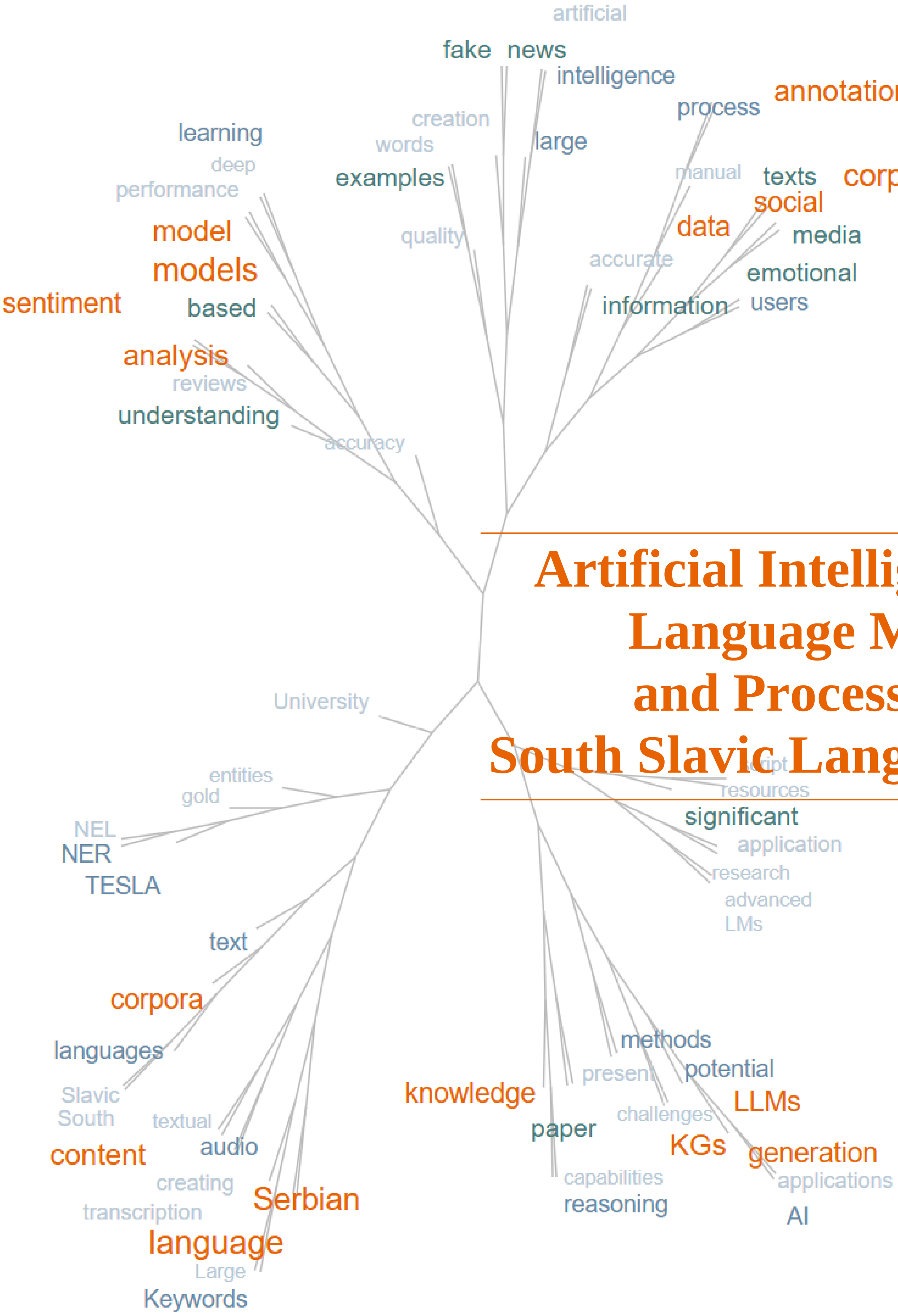
This paper will demonstrate how the analysis of dictionary definitions (specifically, definitions from the first Croatian web dictionary *Mrežnik*) reveals lexical-syntactic patterns that indicate hyponymy and meronymy, as well as their subtypes (for meronymy, for example, relations like object > functional component, group > individual, mixture > ingredient, etc.). It will show how these identified patterns are then tested in the Croatian web corpus (*hrWaC*) and later in the specialized knowledge corpus (*Croatian Linguistic Corpus*). The dictionary and corpus data are analyzed using Python and CQL, leading not only to pattern identification but also to numerous examples of hyponyms and hypernyms, as well as meronyms and holonyms, providing fertile ground for linguistic research. Consequently, the paper will briefly present key

conclusions in the field of lexical semantics, such as the differences between hyponymy and meronymy, their intersections, quasi-relations, autohyponymy, and automeronymy.

Research into lexical-semantic relations based on lexical-syntactic patterns is more common in foreign linguistics, and the results show that such research needs to be adapted to inflectional languages like Croatian (and other South Slavic languages), particularly due to categories such as case, verb valency, prepositional-case expressions, etc.

Keywords: *hyponymy, meronymy, lexical-syntactic pattern, dictionary, dictionary definition, Croatian Web Dictionary – Mrežnik*

Artificial Intelligence, Language Models and Processing of South Slavic Languages



Saša Petalinkar
University in Belgrade, Serbia
E-mail: sasa5linkar@gmail.com

Milica Ikonić Nešić
Faculty of Philology, University in Belgrade, Serbia
E-mail: milica.ikonic.nesic@fil.bg.ac

Automating Synset Example Creation: A Case Study with Serbian Wordnet and ChatGPT

Wordnets are lexical resources that organize words by their meanings, facilitating semantic understanding. The Serbian Wordnet, developed as part of the Balkanet project, serves as a crucial tool for linguistic research and applications. An essential component for human comprehension of synsets—groups of synonymous words—are illustrative examples. However, finding examples that precisely match the exact meaning of words within a synset is a time-consuming and challenging task. This paper proposes an innovative solution to this problem by utilizing a Language Model (LM) to generate synthetic examples for each word in a synset, based on the provided definitions. Our approach leverages the capabilities of advanced LMs, specifically ChatGPT, to create contextually appropriate and semantically accurate examples. This method aims to enhance the usability and accessibility of wordnets by automating the generation of illustrative sentences, thereby saving significant time and effort for researchers and lexicographers.

To evaluate the validity and effectiveness of this solution, we generated synthetic examples for a selection of synsets from the Serbian Wordnet and conducted a meticulous analysis and rating of these examples. The evaluation criteria included semantic accuracy, contextual relevance, and overall coherence of the generated sentences.

Our findings indicate that the LM-generated examples are highly effective in capturing the intended meanings of the words within synsets. The automated process not only produces high-quality examples but also significantly reduces the labor-intensive task of manual example creation. The results demonstrate that LMs, such as ChatGPT, can serve as powerful tools in enhancing lexical resources like wordnets.

This research contributes to the field of computational linguistics by presenting a scalable and efficient method, for example generation in wordnets. It underscores the potential of LMs in supporting linguistic resource development and opens new avenues for their application in various natural language processing tasks. Future work will focus on further refining the generation process, exploring the use of other advanced LMs, and expanding the application to other languages and lexical resources.

In conclusion, the integration of LMs for synthetic example generation in wordnets represents a significant advancement in lexical resource management, offering a promising solution to the challenge of creating illustrative examples for synsets.

Keywords: *LM, Serbian, synset, WordNet, ChatGPT*

Acknowledgment: *This research was supported by the Science Fund of the Republic of Serbia, #7276, Text Embeddings - Serbian Language Applications - TESLA.*

Milica Ikonić Nešić, Miloš Utvić
Faculty of Philology, University of Belgrade, Serbia
E-mail: {milica.ikonic.nesic|milos.utvic}@fil.bg.ac.rs

Overview of the Tesla-Ner-Nel-Gold Dataset: Showcase on Serbian-English Parallel Corpus

As self-explanatory name suggests, TESLA-NER-NEL-gold is designed as an expansion of the existing srpELTeC-gold dataset (Krstev et. al, 2021), to be used as a gold dataset to train models for named entities (NE) recognition (NER) and linking (NEL). This paper gives an overview of the annotation process applied to the TESLA-NER-NEL-gold subset composed of parallel Serbian-English texts from corpus SrpEngKor (Krstev & Vitas 2011)..

There are four levels of annotation in SrpEngKor-TESLA: part-of-speech (PoS), lemma, NE class and NE linking to the open knowledge base Wikidata. The NE classes encompass names of persons, places, organizations, ethnicities (demonyms), events, and works of art. Also, corpus texts as a whole are associated with their bibliographical metadata (title, author etc.) and text register. The SrpEngKor-TESLA statistical description will be provided, detailing the corpus size (number of corpus texts and sentences) and the distribution of PoS and NER classes, and text registers.

The dataset annotation process follows a specific pipeline: it begins with automatic annotation using the SrpCNNeL and Jerteh-355-tesla models, followed by manual correction by annotators using the INCEPTION platform, and final post-correction by a supervisor. Entity linking is also conducted on the INCEPTION platform to link recognized named entities to the appropriate items in Wikidata. Additionally, this paper will present a model for recognition and linking of the named entities with Wikidata, trained on the TESLA-NER-NEL-gold dataset.

Beyond monolingual exploration of the annotated corpora, we will illustrate the benefits of rich annotations on parallel corpora through cross-language queries. This approach provides opportunities to explore, extract, and learn different language patterns.

Keywords: *parallel corpora, named entities, NER, NEL, Serbian, English*

Acknowledgment: *This research was supported by the Science Fund of the Republic of Serbia, #7276, Text Embeddings - Serbian Language Applications - TESLA.*

References

[1] Cvetana Krstev and Branislava Šandrih Todorović and Ranka Stanković and Milica Ikonić Nešić. (2021). SrpELTeC-gold - Named Entity Recognition Training Corpus for Serbian. ELG, https://live.european-language-grid.eu/catalogue/corpus/9485, 1.0.

[2] Cvetana Krstev, Duško Vitas, "An Aligned English-Serbian Corpus", In: ELLSIIR Proceedings (English Language and Literature Studies: Image, Identity, Reality), Volume I, Belgrade, 4-6 December 2009, eds. N. Tomović & J. Vujić, pp. 495-508, Faculty of Philology, University of Belgrade, ISBN 978-86-6153-005-0, 2011.

mr Milena Šošić
doktorand na Matematičkom fakultetu Univerziteta u Beogradu
E-mail: milena.sosic@gmail.com, pd202030@alas.matf.bg.ac.rs

prof. dr Ranka Stanković
vanredni profesor na Rudarsko-Geološkom fakultetu Univerziteta u Beogradu
E-mail: ranka.stankovic@rgf.bg.ac.rs

prof. dr Jelena Graovac
vanredni profesor na Matematičkom fakultetu Univerziteta u Beogradu
E-mail: jgraovac@matf.bg.ac.rs

Social-Emo.Sr: Emotional Multi-Label Categorization of Conversational Messages from Social Networks X and Reddit

In the digital environment of South Slavic languages, emotion analysis in texts on social media is becoming increasingly important for understanding public opinion, creating personalized content, and analyzing user interactions. This presentation presents a detailed methodology and results of corpus annotation in the Serbian language according to Plutchik's categorization model, which identifies eight basic emotional categories: joy, sadness, anger, fear, trust, disgust, anticipation, and surprise. The aim of the research is to analyze the emotional content of texts taken from social media X (formerly Twitter) and Reddit, each collection containing around 17,000 individual messages and approximately 5,000 complete conversations. The corpus annotation process involved several stages: data collection and preparation, manual annotation by experts, verification of annotation accuracy, and statistical analysis of the harmonized labels. By using a multi-label annotation approach, a richer and more qualitative analysis of emotional states was made possible, with particular significance for the application in analyzing complex emotional content found on social media.

To collect data, automated tools were used to download conversations written in Serbian from social media accounts that address current social, political, musical, and sports topics. Data preparation involved additional selection of messages to ensure the quality of their content, while maintaining the conversational structure of the retrieved data. During data preparation, messages were preliminarily annotated using automatic methods, employing both classical and advanced computational linguistics techniques to improve the efficiency of the manual labeling process. Teams of linguists and psychologists reviewed and assessed the automatically assigned labels for their validity concerning the textual content to which they were assigned. To ensure high accuracy and consistency, standardized procedures were used for training annotators and verifying their evaluations through statistical measures of annotation reliability. The analysis of annotation reliability demonstrated that it is possible to classify emotions in texts from social media in Serbian using Plutchik's model. Statistical data analysis revealed significant distributions of emotions in the messages and provided insights into users' emotional reactions to various emotional stimuli and thematic content.

The multi-label categorized emotional corpus in Serbian Social-Emo.SR represents a significant advancement toward a deeper understanding of emotional dynamics on social media among users. In addition to enriching linguistic resources for the Serbian language, this corpus opens new possibilities for application in research, commercial applications, and enhancing mental health analysis of the population. The potential application of modern methodologies on the developed corpus would enable the creation of useful tools for recognizing and reflecting

the complexity of human emotions in the current digital world within the Serbian-speaking community. The corpus will be published under open license CC-BY-4.0.

Keywords: *emotions, Plutchik's model, annotation, corpus, social media, Serbian language*

Acknowledgment: *This research was supported by the Science Fund of the Republic of Serbia, #7276, Text Embeddings - Serbian Language Applications - TESLA.*

Mihailo Škorić
Društvo za jezičke resurse i tehnologije JeRTeh
E-mail: mihailo@jerteh.rs

New Language Models for South Slavic Languages

The report will present the challenges and perspectives of modeling South Slavic languages, especially the general language models built on the transformer architecture (BERT, GPT), available corpora of texts for training those models, and the quantity and quality of those corpora. The presentation will offer an overview of the available data and models, primarily the latest textual corpora. The first corpus, *Umbrella*, represents the umbrella web corpus of South Slavic languages and at the same time the largest corpus of texts in the region, includes all other currently available regional web corpora and contains over eighteen billion words. The second corpus, *S.T.A.R.S*, gathers academic works written in the Serbian language and includes, most notably, eleven thousand dissertations downloaded from the NARDUS platform, and a large number of scientific and professional works downloaded from various open repositories that are included in the *eScience* system. In addition, academic corpora of other South Slavic languages will be discussed, which were created from works stored on various web platforms: DABAR (for the Croatian language), the repositories of the universities in Maribor, Ljubljana, Primorska and Nova Gorica, and the DiRROS and REVIS repositories (for the Slovene language), the repository of the universities in Zenica, Sarajevo and East Sarajevo (for the Bosnian language), the repository of the University of Goce Delčev and St. Kliment Ohridski (for the Macedonian language) and the repository of the University of Montenegro (for Montenegrin). Finally, we will talk about new models for text vectorization in South Slavic languages, which were trained using the aforementioned corpora. An analysis of their performance on a number of previously established tasks will be presented, with reference to the model performance and improvements over models trained on the previous generation of the corpora.

Keywords: *Large text corpora, language models, South Slavic languages*

Acknowledgment: *This research was supported by the Science Fund of the Republic of Serbia, #7276, Text Embeddings - Serbian Language Applications - TESLA.*

Danka Jokić, Ranka Stanković, Jelena Jaćimović
Društvo za jezičke resurse i tehnologije JeRTeh
E-mail: danka.jokic@afrodita.rcub.bg.ac.rs; ranka.stankovic@rgf.bg.ac.rs; jelena.jacimovic@stomf.bg.ac.rs

Knowledge Graphs in the Era of Large Language Models: Opportunities and Challenges

The emergence of large language models (LLMs) has significantly impacted the field of artificial intelligence (AI) by excelling at language processing and generation tasks. However, a critical limitation of LLMs lies in their lack of structured knowledge and reasoning capabilities, hindering their effectiveness in real-world applications that demand factual accuracy and context-aware reasoning. Knowledge graphs (KGs), on the other hand, offer a compelling solution. By representing entities and their relationships in a machine-readable format, KGs provide a rich source of structured knowledge. This convergence presents a unique opportunity to explore their symbiotic relationship: leveraging KGs to empower LLMs for the development of next-generation AI applications.

This synergy has the potential to significantly enhance LLM capabilities in several key areas. First, grounding LLMs in the structured knowledge of KGs can substantially improve their ability to comprehend factual information and generate more accurate and reliable responses to complex questions. In addition, KGs also offer the necessary context and relationships to enable LLMs to perform more sophisticated reasoning tasks and incorporate commonsense knowledge into their reasoning processes, leading to more nuanced and human-like understanding. Furthermore, the explicit relationships within KGs can be harnessed to explain the reasoning behind LLM outputs, directly addressing a critical challenge in interpretable AI.

While the potential is undeniable, challenges require attention. Developing effective methods for LLMs to jointly learn from text data and knowledge graphs is crucial for successful integration. Additionally, ensuring the consistency and quality of knowledge graph data is essential, as incomplete or inaccurate information can lead to biased or erroneous LLM results.

Despite these challenges, the integration of LLMs and KGs holds immense potential to revolutionize various AI applications. LLMs empowered by KGs can provide more accurate and comprehensive answers in question-answering systems. Intelligent assistants integrated with KGs can understand and respond to user queries with greater context and factual grounding. Additionally, the combination of LLMs and KGs can lead to the development of more factually accurate and contextually relevant natural language generation, and the creation of more sophisticated and dynamic knowledge representation systems.

Exploring this synergistic relationship between LLMs and KGs is crucial for advancing AI. By addressing the challenges and pursuing effective integration methods, we can pave the way for the development of next-generation AI applications characterized by enhanced understanding, reasoning, and knowledge representation capabilities. In this paper we will present how knowledge graphs can be used to enhance reasoning capabilities of large language models concerning online safety and moderation of harmful textual content in Serbian.

Keywords: *Knowledge Graphs (KGs), Large Language Models (LLMs), Natural Language Processing (NLP), Structured Knowledge, Data Quality, Explainable Artificial Intelligence (EAI), Online Content Safety*

Acknowledgment: *This research was supported by the Science Fund of the Republic of Serbia, #7276, Text Embeddings - Serbian Language Applications - TESLA.*

References

[1] Agrawal, G., Kumarage, T., Alghamdi, Z., & Liu, H. (2024). Can Knowledge Graphs Reduce Hallucinations in LLMs? : A Survey. In K. Duh, H. Gomez, & S. Bethard (Eds.), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) (pp. 3947–3960). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.219.

[2] Kau, A., He, X., Nambissan, A., Astudillo, A., Yin, H., & Aryani, A. (2024). Combining Knowledge Graphs and Large Language Models (arXiv:2407.06564). arXiv. https://doi.org/10.48550/arXiv.2407.06564.

[3] Lin, J. (2022). Leveraging World Knowledge in Implicit Hate Speech Detection. In L. Biester, D. Demszky, Z. Jin, M. Sachan, J. Tetreault, S. Wilson, L. Xiao, & J. Zhao (Eds.), Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI) (pp. 31–39). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.nlp4pi-1.4.

[4] Pan, J., Razniewski, S., Kalo, J. C., Singhania, S., Chen, J., Dietze, S., ... & Graux, D. (2023). Large Language Models and Knowledge Graphs: Opportunities and Challenges. Transactions on Graph Data and Knowledge.

[5] Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J., & Wu, X. (2024). Unifying Large Language Models and Knowledge Graphs: A Roadmap. IEEE Transactions on Knowledge and Data Engineering, 36(7), 3580–3599. IEEE Transactions on Knowledge and Data Engineering. https://doi.org/10.1109/TKDE.2024.3352100.

Nikola Janković
Faculty of Philology, University of Belgrade
E-mail: nikolajankovickv@gmail.com

Jovana Ivaniš
Institute for Serbian Language SANU
E-mail: jovana.ivanis@gmail.com

Usage of the Whisper Large V3 Sr Model for the Transcription of Serbian Spoken Language in Python Programming Language on the Google Colab Platform

This paper presents a *Python* script on the *Google Colab* platform, which uses a fine-tuned model for the transcription of speech in Serbian, *Whisper Large v3 Sr* (https://huggingface.co/Sagicc/whisper-large-v3-sr-cmb), which enables free, high-quality and simple transcription of Serbian speech into text. The motivation for creating this script came from the lack of available tools which would allow researchers to use this model in a straightforward way, without the need for advanced technical knowledge and significant computing resources. This script provides a simple method for uploading audio files, transcribing them using the *Whisper Large V3 Sr model*, and downloading the transcription of the files in a textual format. Firstly, the paper briefly describes the aforementioned model used by the script, followed by a detailed description of how the script functions, along with a user manual. The paper will also present problems which we needed to overcome in order to successfully implement this approach, such as the need for automatic audio file segmentation, the determination of the optimal segmentation parameters, and the implementation of support for different audio file formats. Furthermore, several approaches related to the reduction in the number of errors in the transcription will be presented. We believe that this tool can be of significant importance to researchers, considering that it speeds up the processing of audio data, enables users to process vast amounts of audio material in a short period of time, and provides a consistent and repeatable transcription method, which is very significant for the scientific methodology and the repeatability of language-related research. We also believe that the tool can be useful to other users, considering the fact that it enables creating subtitles for video content, converting audio notes into text, creating automatically generated captions for the deaf and hard of hearing, as well creating textual archives of audio content.

Keywords: *speech transcription, Serbian language, Python, Whisper Large v3, Google Colab, NLP*

Ana Kovačević
Faculty of Security Studies, Belgrade University
E-mail: kana@fb.bg.ac.rs

Fake News and Generative Artificial Intelligence: Risks and Potential Solutions

The fake news phenomenon has been present since the very dawn of human communication, with repeated exposure to false information often resulting in its acceptance as truth. In contemporary society, however, fake news has become extremely dangerous due to the simple, inexpensive, and convincing way it can be created using generative artificial intelligence, and large language models in particular. These models, which continue to make significant advancements, enable the generation of large volumes of content which appear credible, even in the Serbian language. In addition, such models facilitate the creation of personalised content tailored to certain groups or individuals, citing seemingly credible sources, and thus serving to further strengthen the impact of fake news. In a world inundated with information, this problem has become even more pronounced. One of the key channels for the spread of fake news is social media, whose primary goal is to capture and maintain their users' attention with content which reinforces their existing beliefs. This process results in an echo chamber effect, where users receive information which in turn reinforces their preconceptions, making it harder to recognise false information. A secondary, but equally concerning issue caused by fake news is the creation of distrust and confusion among users, ultimately leading to scepticism towards even accurate information. The vast amount of data on social media highlights the need for the application of artificial intelligence, not only to deliver relevant content, but also to detect fake news. Although AI-based approaches can be used to identify false information, questions remain as to the transparency and reliability of these algorithms in carrying out this task. This paper will present the potential uses of artificial intelligence in the creation of fake news, as well as possible solutions to address this problem.

Keywords: *artificial intelligence, large language models, fake news, the risks of artificial intelligence*

Acknowledgment: *The work was supported by the Science Fund of the Republic of Serbia within the framework of the 'IDEAS' Program – Management of New Security Risks Research and Simulation Development, NEWSiMR&D, #7749151*

Maram Alharbi
Lancaster University, Jazan University
E-mail: m.i.alharbi@lancaster.ac.uk

Ruslan Mitkov
Lancaster University and University of Alicante
E-mail: r.mitkov@lancaster.ac.uk

Comparing Rule-Based and Deep Learning Approaches for Understanding Sentiments

In the rapidly evolving domain of sentiment analysis within the hospitality sector, accurately understanding customer sentiment from hotel reviews is increasingly crucial for market intelligence and improving service quality. This study focuses on categorising hotel reviews into positive and negative sentiments using a range of sentiment analysis models. We examine both rule-based approaches—such as VADER, AFINN, and TextBlob—and deep learning methods, including the T5 transformer model and BiLSTM. While rule-based models employ predefined sentiment lexicons, offering simplicity and interpretability, they often fall short in capturing the complexity of emotional nuances and context-dependent sentiments. Conversely, deep learning models, particularly those leveraging advancements in transfer learning, provide a more sophisticated understanding of linguistic intricacies, enabling more effective sentiment detection, even in complex reviews. Our experiments show that rule-based models achieve moderate performance, with F1 scores of 0.79 for TextBlob, 0.77 for AFINN, and 0.78 for VADER. While efficient, these methods often struggle with the contextual subtleties of sentiment. In contrast, deep learning models demonstrated superior performance, with the T5 transformer model achieving an F1 score of 0.96 and the BiLSTM model reaching 0.93. This highlights the potential of deep learning techniques for sentiment analysis in customer reviews, offering more accurate and nuanced sentiment classification than traditional methods. We compare these approaches across two datasets of hotel reviews, assessing the impact of various preprocessing techniques and sentiment analysis models in terms of accuracy, precision, and recall. Our findings emphasise the superior performance of deep learning models, particularly the T5 transformer, in accurately classifying sentiments and addressing the challenges posed by mixed sentiments within reviews. Nonetheless, rule-based models retain their utility in scenarios where computational efficiency is prioritised. This study provides a comprehensive evaluation of sentiment analysis methods within the hospitality industry, offering insights that can enhance both customer experience and business strategy. By comparing traditional and advanced sentiment analysis techniques, we contribute to a deeper understanding of model performance and their practical applicability in real-world settings.

Keywords: *Sentiment Analysis, Rule-based, Deep-learning, Transformers, BiLSTM, T5, VADER, TextBlob, AFINN*

References

[1] R. Mitkov, The Oxford Handbook of Computational Linguistics 2nd edition, Oxford University Press, 2022.

[2] E. Demir, M. Bilgin, Sentiment analysis from Turkish news texts with Bert-based language models and machine learning algorithms, in: 2023 8th International Conference on Computer Science and Engineering (UBMK), IEEE, 2023, pp. 01–04.

[3] A. Ameur, S. Hamdi, S. Ben Yahia, Sentiment analysis for hotel reviews: A systematic literature review, ACM Comput. Surv. 56 (2023). URL: https://doi.org/10.1145/3605152. doi:10.1145/3605152.

[4] S. Mutmainah, D. H. Fudholi, Leveraging Bilstm and Lda for analyzing and dashboarding user feedback in applications, JURNAL MEDIA INFORMATIKA BUDIDARMA 8 (2024) 51–61.

[5] C. Hutto, E. Gilbert, Vader: A parsimonious rule- based model for sentiment analysis of social media text, in: Proceedings of the international AAAI conference on web and social media, volume 8, 2014, pp. 216–225.

[6] J.-P. Colson, Multi-word units in machine translation: why the tip of the iceberg remains problematic–and a tentative corpus-driven solu- tion, Computational and Corpus-based Phraseology (2019) 145.

[7] Mitkov, Ruslan, Computer vs. human intelligence. keynote speech at the Refinitiv conference, City of London., 2019.

[8] R. Stuckardt, Machine-learning-based vs. manually designed approaches to anaphor resolution the best of two worlds. proceedings of the dis- course anaphora and anaphora resolution colloquium, daarc'4, 211-216. Lisbon, Portugal., 2002.

[9] R. Stuckardt, Three algorithms for competence- oriented anaphor resolution. proceedings of the discourse anaphora and anaphora resolution colloquium, daarc'5, 157-163. Sao Miguel, Portugal., 2003.

[10] R.Stuckardt, A machine learning approach to preference strategies for anaphor resolution. in Antonio Branco, Tony McEnery, and Ruslan Mitkov (eds.), anaphora processing: Linguistic, cognitive and computational modelling, 47-72. John Benjamins, Amsterdam/ Philadelphia, 2004.

[11] G. Sreenivas, K. M. Murthy, K. Prit Gopali, N. Eedula, M. H R, Sentiment analysis of hotel reviews - a comparative study, in: 2023 IEEE 8th International Conference for Convergence in Technology (I2CT), 2023, pp. 1–9. doi:10.1109/I2CT57861.2023.10126445.

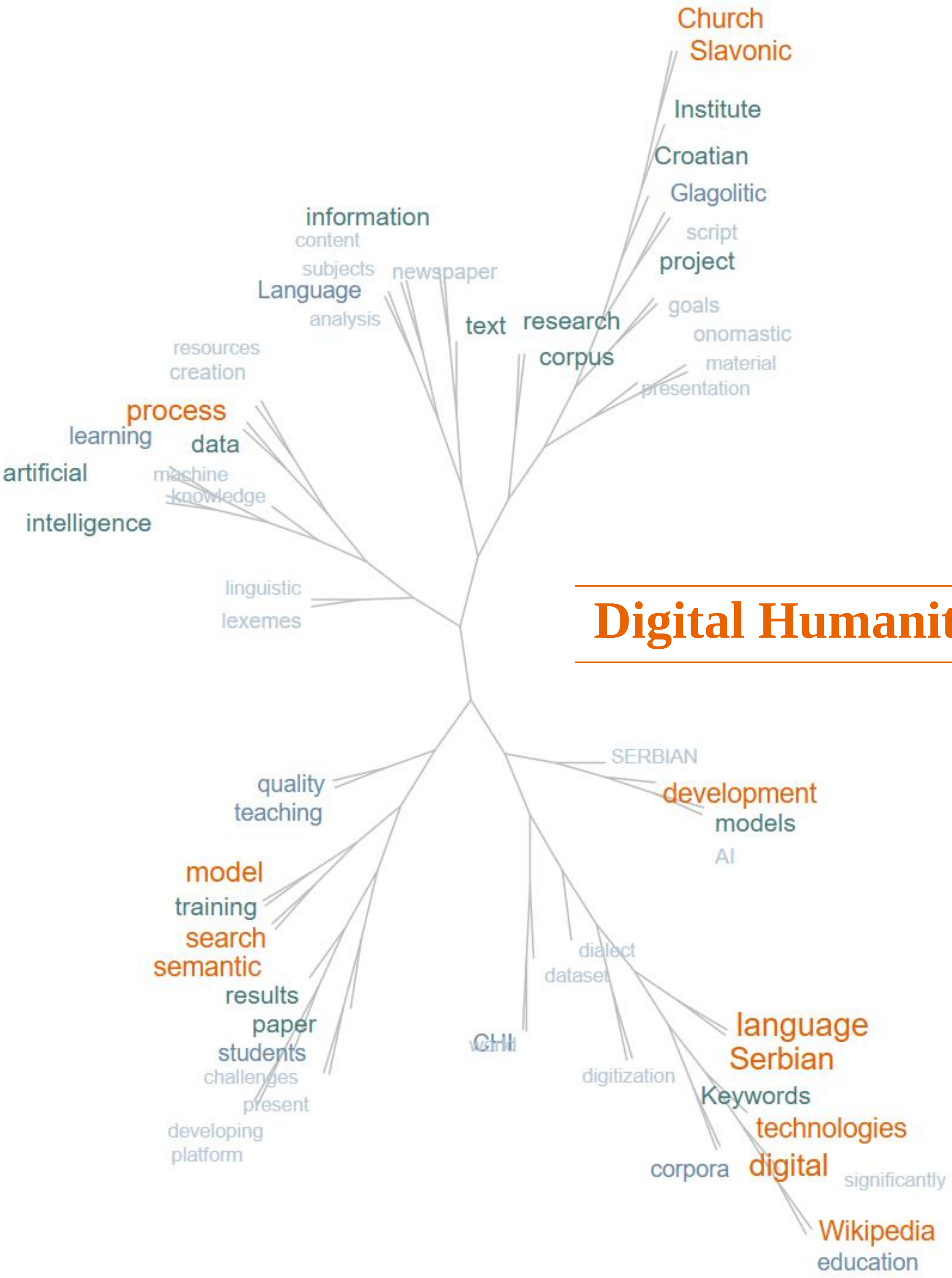
[12] M.R. Orasan, C., Recent Developments in Natural Language Processing. In R. Mitkov (Ed.), The Ox- ford Handbook of Computational Linguistics 2nd edition. Oxford University Press., Oxford Univer- sity Press, 2021.

[13] C. Raffel, N.Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, CoRR abs/1910.10683 (2019). URL: http: //arxiv.org/abs/1910.10683. arXiv:1910.10683.

[14] K. Pipalia, R. Bhadja, M. Shukla, Comparative analysis of different transformer based architectures used in sentiment analysis, in: 2020 9th International Conference System Modelling and Advancement in Research Trends (SMART), 2020, pp. 411–415. doi:10.1109/SMART50582.2020.9337081.

[15] A. Pathak, et al., Comparative analysis of trans- former based language models, in: CS & IT Conference Proceedings, volume 11, CS & IT Conference Proceedings, 2021.

Digital Humanities



Milena M. Stojanović
Faculty of Philology, University of Belgrade
E-mail: stojanovic.milena77@gmail.com

Meanings and Position of Lexemes in The Sphere of Artificial Intelligence in the Serbian Lexis and Perception of New Technologies – Challenges of Contemporary Time

In this paper, four lexemes from the artificial intelligence (*webinar*, *troll*, *phishing* and *chat*) have been analysed from the semantics and linguistic culture point of view. These lexemes are not part of the *Dictionary of the Serbian Language* of Matica srpska. We wanted to find out what attitude students and pupils have towards new technologies and how they know the semantics of these lexemes. The justification for the linguistic cultural approach to the topic is seen in the fact that the essence of language is anthropocentrism, which is the starting point of this research. The semantic approach is seen in the fact that the meaning of lexemes is often upgraded in a certain context, and that terminology is a process of intellectualization of language, and determinology is a form of democratization of language. Through an online questionnaire, divided into four units, the results showed the subjects attitudes regarding the topic through interpretive analysis, systematization, description and interpretation. Structurally, the questions were about the importance of artificial intelligence in the learning process, how familiar are the meanings of terms in artificial intelligence, whether there have prejudices and stereotypes about the new technologies in life and in teaching, and questions related to attitudes about language. The questionnaire was completed by a total of 173 subjects which 72.3% were women and 27.7% were men. Most of the them were from Belgrade 132 (76.3%), from Vojvodina 12 (6.9%), Western Serbia 11 (6.4%), and the least from Eastern Serbia 3 (1.7%). Most of the subjects were high school students – 63%, and students were 38.2%. The results showed that the semantics of the analysed lexemes show a high level of competence, but also go into the domain of pragmatics and linguistic culturology. The analysis confirmed that terminologization is the process of intellectualization of language, and determinologization is the process of democratization, and that in this way the lexical fund is significantly enriched. The results of the role of artificial intelligence in the learning process show that the knowledge-acquiring population has a polarized attitude towards the use of artificial intelligence in the learning process, and they were explicit that Chat GPT, as an additional tool to traditional teaching, is not a good idea (no = 73, and yes = 45, I don't know = 57). The results show that modern technologies are criticized for potential misuse and often questionable quality of information placed mostly, and that professors would not improve the quality of teaching if they use modern technologies.

Keywords: *artificial intelligence, Chat GPT, modern technologies, linguistic culturology, stereotype, prejudice, digital environment*

Nebojša Ratković
Vikimedija Srbije, Menadžer obrazovnog programa
E-mail: nebojsa.ratkovic@vikimedija.org

Integration of Serbian Language Version of Wikipedia into Educational Systems and the Advancement of Language Technologies

The Serbian-language version of Wikipedia stands as one of the most comprehensive and accessible sources of knowledge in the digital era. Integrating Serbian-language Wikipedia into educational systems can significantly contribute to modernizing education, enhancing language and technological literacy, and fostering the development of language technologies among students. This paper explores the opportunities and challenges associated with using Wikipedia in teaching, with a particular emphasis on the creation of digital corpora and the advancement of language technologies. It also considers strategic steps for integrating Wikipedia into national educational systems, offering recommendations for education policy that would facilitate the systematic use of Wikipedia in schools and universities, including teacher training, curriculum development, and the creation of digital platforms to support this integration.

The first part of the paper examines Wikipedia's role as a teaching tool and its potential to enhance teaching materials and methodologies. It also explores how Wikipedia can be used as a platform for developing students; critical thinking, research skills, and digital literacy. To illustrate this, the paper presents cases from schools and universities where Wikipedia has been effectively integrated into the curriculum, along with feedback from teachers and students regarding this integration.

The second part of the paper delves into the technical aspects of integrating Wikipedia in education. It explores how Wikipedia articles can be used to create digital corpora that serve as the foundation for the development and enhancement of language technologies. The use of Wikipedia in education contributes significantly to the construction of digital corpora by enriching and standardizing language content, which enables better analysis and the development of language technologies, such as natural language processing and machine translation tools. This process encourages the systematic digitization of languages and the creation of resources that are vital for the continuous progress and implementation of language technologies in various fields.

Keywords: *Wikipedia, education. digital corpora, language technologies*

Ana Mihaljević
Staroslavenski institut, Zagreb, Hrvatska
E-mail: amihaljevic@stin.hr

Croatian Church Slavonic Language and Glagolitic Script in the Digital Environment

As of January 1, 2024, at the Old Church Slavonic Institute in Zagreb the project Development of the Digital Infrastructure Model of the Old Church Slavonic Institute – DigiSTIN has started as one of the NextGenerationEU projects funded by the European Union. The goal of the project

is to design a model for developing the digital infrastructure of the Old Church Slavonic Institute, which will then be applied in practice. The project primarily focuses on developing the Institute's website stin.hr, aiming for it to not only provide information about the organization and activities of the Old Church Slavonic Institute but also to serve as a hub for information on the Croatian Church Slavonic language, as well as the Croatian Glagolitic script and Glagolitism.

One of the main goals of the project is to create an online version of the Dictionary of the Croatian Redaction of Church Slavonic, which has so far only been published in print. The compilation of this dictionary is the core activity of the Old Church Slavonic Institute. The presentation will discuss the current state of its digitalization. Alongside the digitalization of the Dictionary, the corpus for its compilation is also being digitized (previously, the catalogs were digitized, and now the comparative Croatian Church Slavonic – Latin – Greek glossaries are being digitized), and a searchable reverse dictionary is compiled. An e-grammar of the Croatian Church Slavonic language, based on the Institute's printed grammar Hrvatski crkvenoslavenski jezik, is also developed.

As part of the DigiSTIN project, the expansion of the Croatian Church Slavonic language corpus available in the beram.stin.hr database is planned. Manuscripts and printed sources are currently being scanned and converted into machine-readable text format (OCR and HTR). To improve the quality of transcriptions and the speed of the process, new models for reading manuscripts are being developed within the Transkribus platform, tailored to specific forms of the Glagolitic script.

Subpages have been opened on the Old Church Slavonic Institute website, including Glagoljica u školi (Glagolitic Script in Schools), intended for popularizing the Glagolitic script among school children, where content designed for use in Croatian Language and other school subjects is published, as well as a subpage with games aimed at learning the Glagolitic script, as well as the vocabulary and grammar of the Croatian Church Slavonic language. The presentation will introduce the starting points and goals of the project, what has been achieved so far, and the plans for the future.

Keywords: *Glagolitic script, Croatian Church Slavonic language, Development of the Digital Infrastructure Model of the Old Church Slavonic Institute – DigiSTIN, online dictionary, educational games, corpus, popularization*

Miloš Košprdić, Gorana Gojić, Adela Ljajić, Dragiša Mišković
Istraživačko-razvojni institut za veštačku inteligenciju Srbije
E-mail: {milos.kosprdic|gorana.gojic|adela.ljajic|dragisa.miskovic}@ivi.ac.rs

Development of a Semantic Search Model for Serbian

The development of large language models represents a significant advancement in natural language processing, enabling efficient semantic search and text understanding. This paper presents the training process of a large language model for semantic search in Serbian, focusing on the task of passage ranking. The model is based on the msmaarco-bert-base-dot-v5 architecture and adapted for asymmetric semantic search.

To facilitate model training in Serbian, we used the MSMarco dataset, which was automatically translated from English to Serbian using Google Translate. This dataset encompasses a wide range of questions and answers, allowing the model to learn the richness of semantic connections in Serbian.

The aim of this paper is to showcase the current results in developing a semantic search model for Serbian and to demonstrate the feasibility of successfully training a semantic search model in a language with limited resources. In addition to the technical aspects of training, we present the challenges encountered during the data translation process and the strategies employed to overcome them.

The model training process was conducted over 60 epochs with a batch size of 64, using a single 40GB A100 GPU on the National Platform for Artificial Intelligence in Kragujevac. The model was tested on a subset of the MSMarco test set for queries and answers, achieving satisfactory performance in terms of search accuracy and relevance, as shown in Table 1.

Evaluation Results of the Model in English and Serbian

model	**Acc@10**	**P@10**	**R@10**	**MRR@10**	**NDCG@10**	**MAP@100**
English	68.51	6.91	68.10	0.3700	0.4435	0.3802
Serbian	54.03	5.44	53.7	0.2923	0.3501	0.3024

Acc@10 – Accuracy, P@10 – Precision, R@10 – Recall, MRR@10 – Mean Reciprocal Rank, NDGC@10 – Normalized Discounted Cumulative Gain, MAP@100 – Mean Average Precision.

Our results indicate that the model can satisfactorily rank passages in Serbian, demonstrating the robustness and adaptability of the msmaarco-bert-base-dot-v5 architecture. This model contributes to the development of NLP tools for Serbian, opening new possibilities for the application of semantic search in various domains.

This work contributes to the development of language models for less-represented languages, providing a framework and methodology applicable to other languages facing similar preprocessing and training challenges. Our results confirm that automatic data translation, with certain corrections, can be an effective approach for training high-quality language models, thereby enhancing support for Serbian in the digital environment. Future work will focus on improving the model through further fine-tuning and evaluation on specific semantic search tasks, as well as its application in real-world search systems.

Keywords: *Semantic Search, Passage Ranking Task, MSMarco dataset, LLM, Serbian*

Dr. Snežana Petrović
Institute of Serbian Language SANU
E-mail: snezzanaa@gmail.com

Dr. Mirjana Petrović-Savić
Institute of the Serbian Language SANU
E-mail: mirjana.petrovic@isj.sanu.ac.rs

Dr. Ana Španović
Institute of the Serbian Language SANU
E-mail: tesicana@gmail.com

MSc Lenka Bajčetić
Innovation Center of the Faculty of Electrical Engineering in Belgrade d.o.o.
E-mail: lenka.bajcetic@gmail.com

MSc Matija Nešović
Institute of Serbian Language SANU
E-mail: nesovic1998@gmail.com

MSc Jovana Todorić
Institute of Serbian Language SANU
E-mail: jovanatodoric080@gmail.com

Digitization of the Sanu Onomastic Committee's Onomastic Material—Importance, Goals and First Steps

This paper will describe the importance, goals, and first steps in digitizing the SANU Onomastic Committee's onomastic material, which includes about 750,000 onomastic slips collected from the entire Serbian language area from 1975 to the present day.

The description of the first steps in the digitization of that material includes the presentation of the preparation and process of scanning and documenting, then the modelling of various data recorded on the onomastic material sheets to create a structured database (Figure 1) and the creation of a user interface for accessing the database and entering data (Figure 2). The process of creating and implementing the interface will be presented, with a review of the problems and doubts that arose during the creation process and the implemented solutions.

The ultimate goal of digitizing onomastic material is to publish it on a multi-searchable, open-type platform. This platform is designed to be accessible to everyone, fostering a sense of inclusivity and community among researchers and the general public. It will allow onomastic data to be organized, linked, and displayed in multiple ways, including various types of searches and data visualization in the form of digital maps.

Keywords: *Serbian language, digitization, onomastics, interface development, data visualization.*

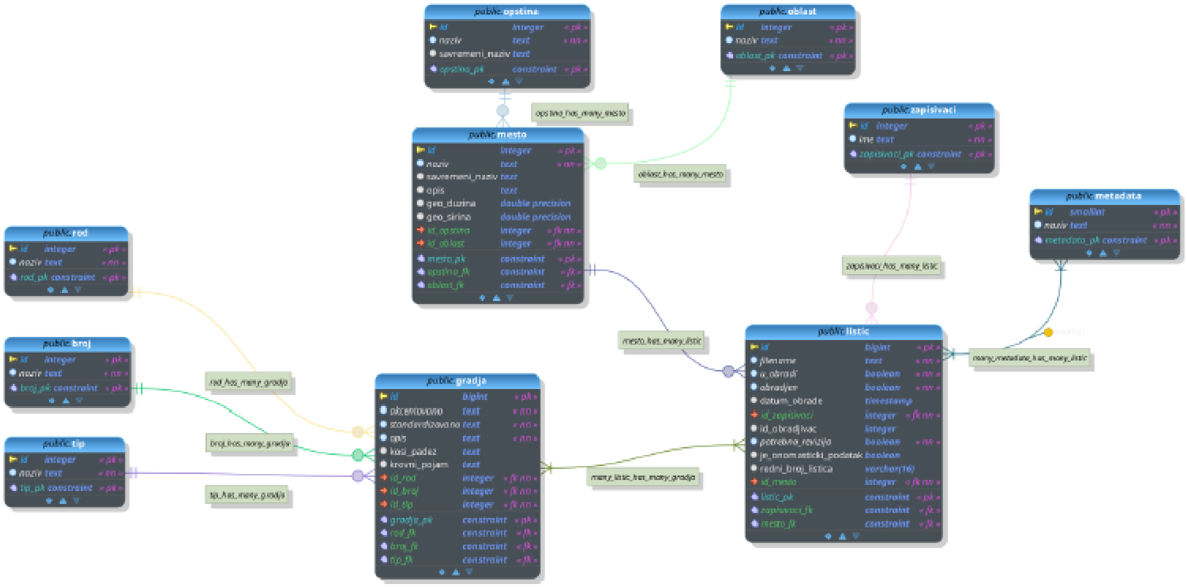
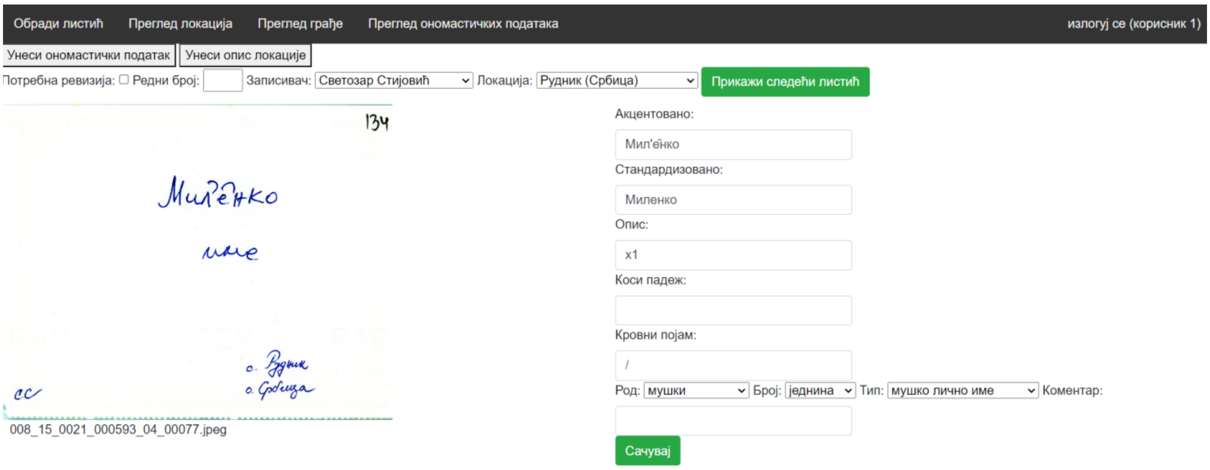


Figure 1. Schema of the onomastic structure model



Потребна ревизија: □ Редни број: Записивач: Светозар Стијовић Локација: Рудник (Србица)

134

Мил'е̑нко

име

с. Рудник
о. Србица

сс

008_15_0021_000593_04_00077.jpeg

Акцентовано: Мил'е̑нко

Стандардизовано: Миленко

Опис: x1

Коси падеж:

Кровни појам: /

Род: мушки Број: једнина Тип: мушко лично име Коментар:

Figure 2. Interface for entering onomastic data

dr Vasilije Milnović, dr Aleksandra Trtovac
Univerzitetska biblioteka „Svetozar Marković“, Beograd
E-mail: {milnovic|aleksandra}@unilib.rs

prof. dr Cvetana Krstev, prof. dr Ranka Stanković, prof. dr Duško Vitas
Društvo za jezičke resurse i tehnologije - JeRTeh
E-mail: {cvetana|ranka|dusko}@jerteh.rs

Improving Machine Understanding of Text and Finding Information in Historical Newspapers in Serbia

The development and application of advanced language models and technologies can significantly improve the accuracy of finding information in the text of scanned newspaper pages and linking them to knowledge bases available on the Internet. Given the increasing attention paid to artificial intelligence in digital humanities and archival research, there is a need for the development of Archival Linked Data. Although artificial intelligence techniques in research and practical solutions have much more to offer, they offer great potential for digital archives to increase the volume of linked data and innovate the way to access that data. Automating the process and reducing the resources required to produce linked archival data enables more efficient work by experts, while it is crucial that any implementation is focused not only on its use, but also on its limitations, biases and ethical implications.

The collection of *Historical Newspapers* of the "Svetozar Marković" University Library represents a huge and significant resource that was the basis for several research and library-archival projects. To modernize the access to these text collections, the new research relied on the activities of earlier, successfully implemented, projects related to distant reading (*Distant Reading*). This paper will present several innovations, i.e. segments of improving the semantic visibility of historical newspapers:

1. Improving the search by expanding the query to include the search grammatical forms of key word, relying on the web services of the Society for Language Resources and Technologies and electronic dictionaries of the Serbian language.
2. Finding, tagging and extracting key information, such as titles, dates, people, organizations and locations.
3. Identification of important entities and relations between them, using semantic networks such as WordNet, Wikidata, GeoNames, as well as vector representations of words developed within the TESLA (Text Embeddings - Serbian Language Applications) project.
4. Using location information to display on a map the location mentioned in newspaper articles, associating geographic locations with specific events or time points.
5. Development of interactive visualizations that allow users to explore and understand historical information from newspaper articles, using graphs, maps and other visual elements.

The results that will be presented should enable improved text search and information extraction from scanned newspaper pages, visualizations that enable the research of historical topics and events, effective search and recommendations to users based on the analysis of newspaper content, as well as the education of librarians and researchers through prepared case studies.

Keywords: *language models, Serbian language, digital humanities, Archival Linked Data, artificial intelligence, Distant Reading, knowledge graphs*

Andrija Sagić
“Milutin Bojić” Library, Belgrade, Head of Digital Development Department,
E-mail: andrija.sagic@milutinbojic.org.rs

Cultural Heritage Institutions in the Era of AI

Cultural Heritage Institutions (CHI) have a large and valuable source of materials (text and images) in their funds that can be used in machine learning. Serbian is classified as Low Resource Language. The most available content is generated from web sources, but the relevance and quality is questionable for training an AI model. The resources are often fake news, propaganda and click catchers.

The initiative "Collections as Data", started in 2016, see the value of CHI digital collections in a dataset format. Digitization and preparation of collections for AI models development, today, can be an important segment in CHI. How, in a best way, can we use those digital collections? How to prepare them? What licence to assign? There are the questions that raise and need to be answered with cooperation with ML development community.

What are the ways that CHI can participate to get better development of low resource AI models, for example text, image and audio generation, speech transcription and other?

In this presentation will be presented some possible solutions that include CHI, selected practice around the world comparable with Serbian.

CHI during the years, around the world, have established digital labs that prepare and publish their datasets (from digital collections), tools for preparation and reuse, organize events...

Will the forming of the first CHI digital lab in Serbia, following the practice around the world, can partially solve, the quality resource missing for AI model training? We hope so.

Keywords: *digitization, cultural heritage, machine learning, cultural institutions*

Srdjan Šućur, Jelena Marković
University of East Sarajevo, Faculty of Philosophy Pale, Department of English
E-mail: {srdjan.sucur|jelena.markovic}@ffuis.edu.ba

The Digitalisation of the Serbian Cultural Heritage of the Jekavian Dialect (1840–1920) with the Faculty of Philosophy Pale Digital Humanities Centre (Stage One)

"What's not written down doesn't exist […]," are the words with which Meša Selimović starts "The Fortress." In a digital environment, one of the interpretations of these words could be "What isn't digitised doesn't exist." The significance and role of digital humanities are, among other things, reflected in the conversion of the analog into the digital, i.e., in digital reformatting. This is one of the goals of the project *the Digitalisation of the Serbian Cultural Heritage of the Jekavian Dialect (1840–1920) with the Faculty of Philosophy Pale Digital Humanities Centre (stage one)*.

A corpus of the representative literary works written in the jekavian dialect of the Serbian language, whose formation is one of the goals of the project, presents an addition to the SrpELTeC+ corpus (which is, due to objective reasons, predominantly ekavian), and whose development started within the finished COST action named Distant Reading for European Literary History (CA16204). Given how many native speakers of this dialect there are, the jekavian dialect is underrepresented in the Serbian language corpora currently available, and is consequently unavailable for corpus research, or to a wider readership. Initially, it was devised that this corpus should consist of 30 publications, and the "stage one" in its name suggests that future expansion and additions to the corpus are planned.

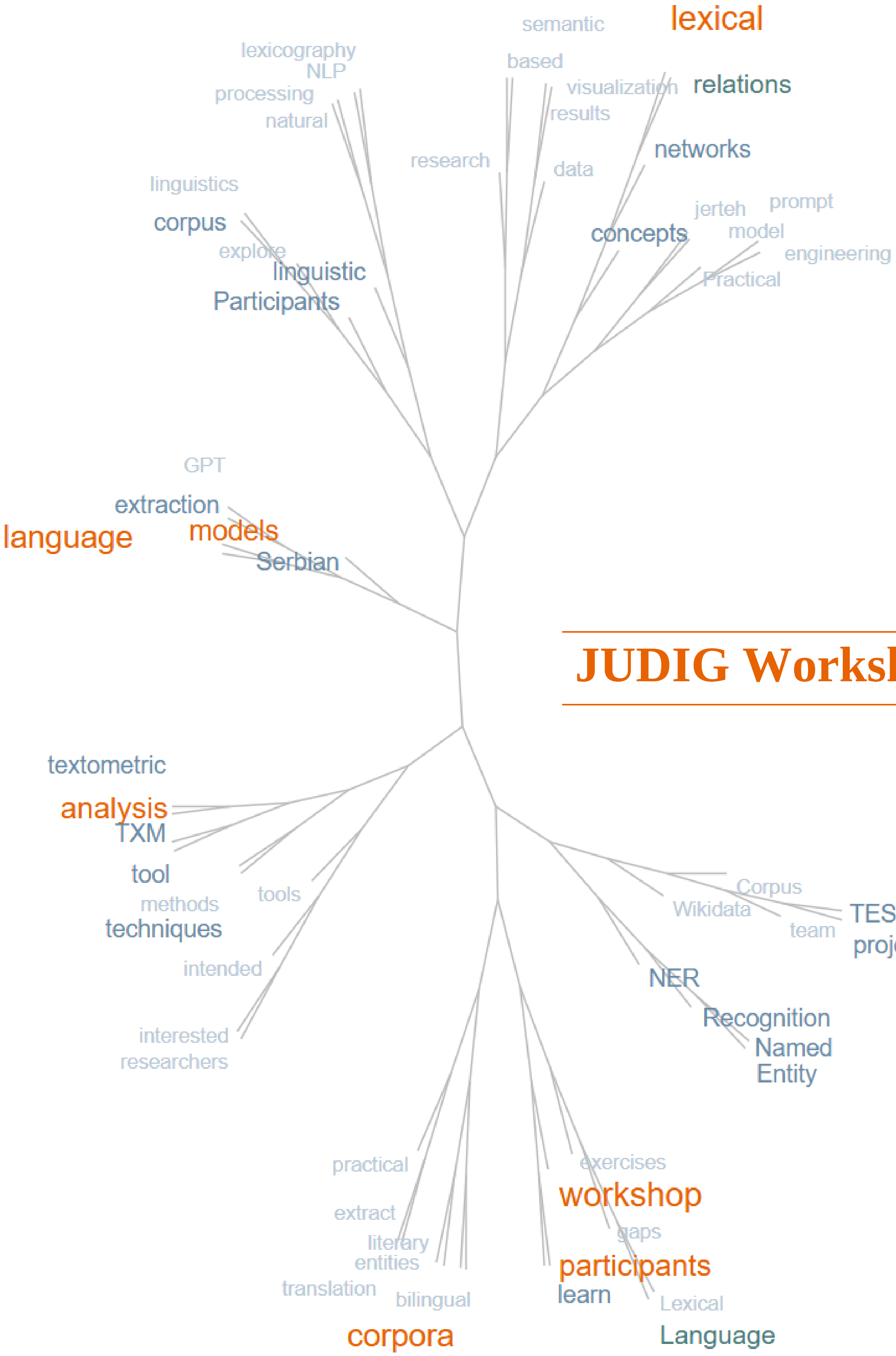
In this project description, we present the current developments and challenges met in the realisation of the 4 planned stages, which are as follows: text selection, workshops organised by the Language Resources and Technologies Society JerTeh, selected texts digitising, publication of the digitised texts on the DHC website, as well as beta testing.

Stage one includes short stories by authors such as: Petar Kočić, Svetozar Ćorović, Vaso Kondić, Marko Popović, Milan Trifunović, Joanikije Pamučina, travelogues by: Konstantin Hadžiristić, Jovo Besarović, Sava Kosanović, Risto Besarović, Sava Pješčić and Marko Car, and novels by: Svetozar Ćorović and Radovan Tunguz-Perović.

A metadata database, containing all the relevant information about the digitised titles has been created, since its existence is one of the preconditions that make corpora usable for scientific purposes.

Keywords: *SrpELTeC+, digitalisation, Serbian literary heritage, jekavian dialect*

JUDIG Workshops



Olja Perisic, PhD

Università di Torino, Dipartimento di Lingue e Letterature straniere e Culture moderne
https://unito.webex.com/meet/olja.perisic

Olja Perišić is an assistant professor of Serbian language at the Department of Foreign Languages and Literature, University of Turin. She received her PhD in Digital Humanities from the University of Genoa in 2020. With over ten years of experience as a literary translator and interpreter between Italian and Serbian, she specializes in the application of corpus and digital technologies to teaching Serbian as a foreign language, translation studies, bilingual lexicography, and contrastive grammar. She is the author of the monograph Il corpus per imparare il serbo. Il futuro dell'apprendimento linguistico (2023).

Corpus Query Language (CQL): Lexical gaps in bilingual corpora

The workshop is designed for educators, translators, and researchers interested in utilizing language corpora for foreign language teaching and translation, regardless of their prior experience with corpus linguistics. Participants will explore the methodological foundations of corpus linguistics and how these can be applied across various research areas.

We will explore cross-linguitic lexical inconsistencies, such as concepts that exist in one language but not in another, shaped by cultural differences and linguistic anisomorphism (including polysemy and lexical gaps). Through hands-on exercises, participants will learn to effectively search parallel corpora, progressing from basic to complex queries, including Named Entity Recognition (toponyms, anthroponyms, etc.).

The second half of the workshop will be dedicated to practical application, where participants will independently extract translation equivalents and named entities from the bilingual corpora of literary texts It-Sr-NER and SerbItaCor3_sr.

TESLA project team: Milica Ikonić Nešić, Mihailo Škorić and Saša Palinkar
Univerzitet u Beogradu – Rudarsko-geološki fakultet i Društvo za jezičke resurse i tehnologije JeRTeh

Named-Entity Recognition (NER) and Linking with Wikidata

The workshop will provide participants with an insight into the concepts and techniques of automatic Named Entity Recognition (NER). Participants will learn how to compare models that identify people, places, and organizations in literary works and link them to the corresponding entities on Wikipedia.

The practical part of the workshop will include the use of tools and models, including those based on the vector representation of words, which are created as part of the TESLA Text Embeddings - Serbian Language Applications project (PRIZMA #7276), funded by the Science Fund of the Republic of Serbia. Tools and services available at https://ners.jerteh.rs/ as well as model jerteh-355-tesla, the INCEPTION tool and Wikidata will be used. Attendees will get to know other resources that are being developed as part of the TESLA project (https://tesla.rgf.bg.ac.rs/).

TESLA project team: Ranka Stanković, Cvetana Krstev and Duško Vitas
Univerzitet u Beogradu – Rudarsko-geološki fakultet i Društvo za jezičke resurse i tehnologije JeRTeh

Corpus analysis: textometry, TXM and other tools

The workshop is intended for everyone who is interested in modern techniques and methods in the processing of natural languages. Participants will first get acquainted with the concept and methods of textometric analysis embedded in the TXM tool, and then with the models and resources developed for the Serbian language by the Society for Language Resources and Technologies JeRTeh.

The goal of the workshop is to show the participants how they can use textometric analysis on ready-made JeRTeh corpora, and then create their own corpora. The second part of the workshop will be devoted to the creation and textometric analysis of own corpora using the TXM tool. Texts from the corpus of Serbian novels (1840–1920) SrpELTeC and a parallel corpus of abstracts from the JuDig conference will be prepared for tutorial. Participants will learn about the resources that are being developed as part of the TESLA project (Text Embeddings – Serbian Language Applications, PRIZMA #7276), which is financed by the Science Fund of the Republic of Serbia (https://tesla.rgf.bg.ac.rs/).

Benedikt Perak, assistant professor

Faculty of Humanities and Social Sciences, University of Rijeka

Dragana Špica, assistant professor

Cultural Studies Department University in Pula, Croatia
https://portal.uniri.hr/Portfelj/1078

Benedikt Perak is an assistant professor at the Faculty of Philosophy of the University of Rijeka, where he teaches courses such as Data Science in Culture, Tools and Methods of Digital Linguistics, and Artificial Intelligence and Communication. He is also the leader of the micro-qualification Language Technologies: Text Analysis and Information Extraction. His work includes the application of large-scale language models (LLM) in the analysis of language networks, as well as the use of advanced digital tools for natural language processing and computational semantics. Perak has extensive research experience in the field of digital linguistics, particularly in the application of models for lexical network analysis and computational language processing. He actively uses computer programs such as Python, Networkx and Neo4j for processing and visualizing language data, creating language corpora, and developing tools for computational linguistics.

Dragana Špica is an assistant professor at the Department of Asian Studies of the Faculty of Philosophy of Jurja Dobrila University in Pula, where she teaches Japanese language and related subjects at the undergraduate study of Japanese language and culture and the graduate study of Japanese studies. She is the author of several scientific papers in the field of linguistics and co-author of the monograph Introduction to the Science of the Japanese Language, the first of its kind in the territory of the former Yugoslavia. She is interested in Japanese phonology, adjectives in Japanese and lexicology. She graduated in Belgrade and received her master's and doctorate in Osaka.

Creating Lexical Networks using Large Language Models (LLM) with a Focus on Synonym extraction

This workshop aims to familiarize participants with the techniques of using large language models (GPT-4) for automated extraction of synonyms and antonyms and building lexical networks. During the workshop, participants will learn how to correctly set prompts for language models, define lexical relationships, and use the results for visualization and analysis of semantic structures.

Main workshop topics:

- Introduction to the concepts of lexical networks and relations (synonymy, antonymy, hierarchical relations)
- Using LLMs to extract lexical relations
- Practical demonstrations of prompt-engineering to obtain precise lexical data
- Analysis of results and visualization of lexical networks using a graph model
- Practical applications in NLP and lexicography

Objectives:

- Increasing accuracy and speed of lexical relations extraction through the use of GPT
- Developing prompt-engineering skills for the needs of lexicographic research
- Creation and evaluation of semantic graphs based on the obtained data

The workshop is intended for researchers and practitioners dealing with linguistic analysis, lexicography, and natural language processing (NLP), but deep technical background is not required.

CIP - Каталогизација у публикацији
Народна библиотека Србије, Београд

811.163'322(048)(0.034.2)
004.8(048)(0.034.2)

INTERNATIONAL Conference South Slavic Languages in the Digital Environment JuDig (2024 ; Beograd)

Book of abstracts [Elektronski izvor] / International Conference South Slavic Languages in the Digital Environment JuDig, International Conference South Slavic Languages in the Digital Environment JuDig, November 21-23. 2024, [Belgrade] ; [organisers University of Belgrade - Faculty of Philology [and] Society for Language Resources and Technologies JeRTeh]. - Belgrade : University, Faculty of Philology, 2024 (Belgrade : University, Faculty of Philology). - 1 elektronski optički disk (CD-ROM) : tekst ; 12 cm

Sistemski zahtevi: Nisu navedeni. - Nasl. sa naslovnog ekrana. - Tiraž 100.

ISBN 978-86-6153-754-7

a) Јужнословенски језици -- Рачунарска лингвистика – Апстракти

COBISS.SR-ID 157072905