

Book of Abstracts of the Second

International Conference

The Final Conference of the TESLA Project



Faculty of Mining and Geology, University of Belgrade
26-28 October 2026

ORGANISERS

University of Belgrade – Faculty of Mining and Geology
Language Resources and Technologies Society – JeRTeh
University of Belgrade – Faculty of Philology
Institute for the Serbian Language of SASA
University Library “Svetozar Marković”

Editors:

Prof. Ranka Stanković  0000-0001-5123-6273 University of Belgrade, Faculty of Mining and Geology and Language Resources and Technologies Society – JeRTeh

Dr Aleksandra Marković  0000-0003-2985-3042 Institute for the Serbian Language of SASA and Language Resources and Technologies Society – JeRTeh

Technical editors:

Prof. Ranka Stanković, Aleksandra Marković, PhD
MSc Jovana Rađenović, MA Marina Bagi

Graphic design:

Anđelka Zečević
Aleksandra Tomašević, PhD

Publisher:

University of Belgrade – Faculty of Mining and Geology

For publisher:

Prof. dr Aleksandar Cvjetić

Print:

University of Belgrade – Faculty of Mining and Geology

Circulation:

100 copies

Publication accessible at:

<https://judig.jerteh.rs/2026/judig-2026-book-of-abstracts.pdf>

Website: <https://judig.jerteh.rs/>

E-mail: judig@jerteh.rs

ISBN: 978-86-7352-432-0

DOI: 10.18485/judig_abs.2026.2



<https://creativecommons.org/licenses/by/4.0/>
Works are licensed under the terms and conditions of the Creative Commons Attribution 4.0 International License.

ACKNOWLEDGEMENTS



Republic of Serbia

MINISTRY OF SCIENCE,
TECHNOLOGICAL DEVELOPMENT AND INNOVATION



Science Fund
of the Republic of Serbia



Тесла

FACTORY

(gold sponsor)



INSTITUTE FOR
STANDARDIZATION
OF SERBIA



Office for Information
Technologies and e-Government



We sincerely thank all sponsors who generously supported JUDIG 2026 financially. Their contributions were invaluable, and we greatly appreciate their commitment to the event's success.

PROGRAMME COMMITTEE AND REVIEWERS TEAM

Co-Chairs:

Prof. Ranka Stanković, University of Belgrade, Faculty of Mining and Geology and Language Resources and Technologies Society – JeRTeh

Aleksandra Marković, PhD, Institute for the Serbian Language of SASA and Language Resources and Technologies Society – JeRTeh

Members:

Prof. Agata Savary, University of Paris-Saclay, France

Aleksandra Trtovac, PhD, University Library “Svetozar Marković”, Serbia

Ana Ostroški Anić, PhD, The Institute for the Croatian Language, Croatia

Balša Stipčević, PhD, Assistant Professor, University of Belgrade, Faculty of Philology, Serbia

Biljana Rujević, PhD, University of Belgrade, Faculty of Mining and Geology and Language Resources and Technologies Society – JeRTeh, Serbia

Prof. Cvetana Krstev, Language Resources and Technologies Society – JeRTeh, Serbia

Danilo Aleksić, PhD, Assistant Professor, University of Belgrade, Faculty of Philology, Serbia

Prof. Dušanka Popović, University of Montenegro, Faculty of Philosophy, Montenegro

Prof. Duška Klikovac, University of Belgrade, Faculty of Philology, Serbia

Prof. Duško Vitas, Language Resources and Technologies Society – JeRTeh, Serbia

Prof. Giouli Paraskevi, Aristotle University of Thessaloniki & Institute for Language and Speech Processing, Athena RC, Greece

Prof. Gordana Pavlović-Lazetić, Language Resources and Technologies Society – JeRTeh, Serbia

Prof. Ivan Obradović, Language Resources and Technologies Society – JeRTeh, Serbia

Prof. Jan Odijk, Institute for Language Sciences, Utrecht University, Netherlands

Prof. Jelena Graovac, University of Belgrade, Faculty of Mathematics, Serbia

Prof. Jelena Jovanović, University of Belgrade, Faculty of Organisational Sciences, Serbia

Prof. Jelena Marković, University of East Sarajevo, Faculty of Philosophy, Republika Srpska

Jovan Čudomirović, PhD, Assistant Professor, University of Belgrade, Faculty of Philology, Serbia

Prof. Katerina Zdravkova, Saints Cyril and Methodius University – Skopje, Faculty of Computer Science and Engineering, Macedonia

Kristina Štrkalj Despot, PhD, The Institute for the Croatian Language, Croatia

Lucija Mandić, PhD, Research Center of the Slovenian Academy of Sciences and Arts, Slovenia

Prof. Marko Robnik Šikonja, University of Ljubljana, Faculty for Computer and Information Science, Slovenia

Mihailo Škorić, PhD, University of Belgrade, Faculty of Mining and Geology, Serbia

Milica Dinić Marinković, PhD, Assistant Professor, University of Belgrade, Faculty of Philology, Serbia

Miloš Utvić, PhD, Assistant Professor, University of Belgrade, Faculty of Philology, Serbia

Prof. Nadežda Stojković, University of Niš, Faculty of Philosophy, Serbia

Nebojša Vasiljević, PhD, Petlja Foundation, Serbia

Prof. Nelda Kote, Polytechnic University of Tirana, Albania

Prof. Nevena Ceković, University of Belgrade, Faculty of Philology, Serbia

Olivera Kitanović, PhD, University of Belgrade, Faculty of Mining and Geology, Serbia

Prof. Olja Perišić, University of Turin, Department of Foreign Languages, Literatures and Modern Cultures, Italy

Petar Božović, PhD, University of Montenegro, Faculty of Philology, Montenegro

Prof. Petja Osenova, Sofia University “St. Kliment Ohridski”, Bulgaria

Prof. Rada Stijović, Language Resources and Technologies Society – JeRTeh, Serbia

Prof. Ruslan Mitkov, University of Alicante, Spain

Saša Marjanović, PhD, Assistant Professor, University of Belgrade,
Faculty of Philology, Serbia

Prof. Saša Moderc, University of Belgrade, Faculty of Philology, Serbia

Simon Krek, PhD, “Jožef Stefan” Institute, Artificial Intelligence
Laboratory & University of Ljubljana, Center for Language
Resources and Technologies, Slovenia

Prof. Sonja Nenezić, University of Montenegro, Faculty of Philology,
Montenegro

Srđan Šućur, PhD, Assistant Professor, University of East Sarajevo,
Faculty of Philosophy, Republic of Srpska

Vasilije Milnović, PhD, University Library "Svetozar Marković", Serbia

Verginica Barbu Mititelu, PhD, Research Institute for Artificial
Intelligence, NLP Group, Romanian Academy, Romania

Prof. Vladan Devedžić, University of Belgrade, Faculty of Organisational
Sciences, Serbia

Prof. Vladimir Polomac, University of Kragujevac, Faculty of Philology
and Arts, Serbia

Prof. Lan Yunshi, East China Normal University, Shanghai

ORGANISING COMMITTEE

Co-Chairs:

Prof. Ranka Stanković, University of Belgrade, Faculty of Mining and Geology and Language Resources and Technologies Society – JeRTeh

Milica Dinić Marinković, PhD, Assistant Professor, University of Belgrade, Faculty of Philology and Language Resources and Technologies Society – JeRTeh

Members:

Balša Stipčević, PhD, Assistant Professor, University of Belgrade, Faculty of Philology

Nevena Ceković, PhD, Associate Professor, University of Belgrade, Faculty of Philology

Miloš Utvić, PhD, Assistant Professor, University of Belgrade, Faculty of Philology

MSc Milica Ikonić Nešić, University of Belgrade, Faculty of Philology

Biljana Rujević, PhD, University of Belgrade, Faculty of Mining and Geology and Language Resources and Technologies Society (JeRTeh)

Mihailo Škorić, PhD, Language Resources and Technologies Society (JeRTeh)

Aleksandra Marković, PhD, Institute for the Serbian Language of SASA and University of Belgrade, Faculty of Mining and Geology

Anđelka Zečević, University of Belgrade, Faculty of Mining and Geology

MSc Jovana Rađenović, University of Belgrade, Faculty of Mining and Geology

Olivera Kitanović, PhD, University of Belgrade, Faculty of Mining and Geology

Marijana Bogdanović, PhD, Institute for the Serbian Language of SASA

MA Marina Bagi, Institute for the Serbian Language of SASA

Table of Contents

Editors' Preface	1
-------------------------------	----------

Invited Speakers

Tiago Timponi Torrent

Curation is also something you need: on the role of Language Resources in AI	5
---	----------

Petya Osenova

PressMint: Towards Interoperable Corpora of Historical Newspapers	7
--	----------

Francesca Frontini

CLARIN and the Emerging European Ecosystem for Language Data and Technologies: The Italian Perspective	10
---	-----------

Workshops

Tiago Timponi Torrent, Ely Matos, Marcelo Viridiano

Building Enriched Frame-Based Resources: The FrameNet Brasil Framework	13
---	-----------

Ivana Filipović Petrović, Nikola Ljubešić, Ivan Porupski, Isabell Furkert, Tanja Samardžić, Kristina Ilić, Robert Sviben

Spoken and web corpora in the analysis of language variation	17
---	-----------

Mihailo Škorić, Biljana Rujević, Milica Dinić Marinković, Ranka Stanković, Saša Petalinkar

TESLA: Language Resources. Use of language resources developed within the framework of the TESLA project with a focus on the TESLA portal	23
--	-----------

*Ranka Stanković, Milica Ikonić Nešić, Jovana Radenović,
Olivera Kitanović*

TESLA: From Text to Knowledge Graphs..... 27

Panel Session

**AI4LANG – Integrating Artificial Intelligence for the
Digital Transformation of Language Teaching 34**

Extended Abstracts

Olja Perišić

**Momo Kapor u Digitalnom Okruženju: Književno
Prevodenje i ChatGPT 37**

Maarten Janssen, Anna Jouravel, Piroska Lendvai

**Combining Scholarly Edition and NLP: Historical
Multilingual Parallel Text Collections on TEITOK..... 49**

Бурђа Перовић

**Језичка анализа Q&A комуникације у јавним
позивима за финансирање научноистраживачких
пројеката 61**

*Снежана Петровић, Владимир Поломац, Ана Шпановић,
Матија Нешовић, Јована Тодорић, Ленка Бајчетић*

**Кориснички интерфејс за унос ономастичких
података: искуства, изазови и унапређење модела..... 74**

Вељко Миљанић, Јован Турањанин, Михаило Шкорић

**Мерење квалитета аутоматског пресловљавања
латинице на ћирилицу 84**

Saša Petalinkar, Nikola Vulović

**A Reproducible Codex Skill Workflow for Serbian
Word Sense Disambiguation with Serbian WordNet 95**

Таня Нейчева

**Графичната организација на текста како аналитичен
објект при изграждането на диахронен корпус..... 108**

Jovana Bazić

Korpus javnog diskursa u italijanskom i srpskom jeziku: struktura, metodologija i istraživački potencijal 120

Irena Srdanović, Benedikt Perak, Dragana Špica

Croatian Tour-Guiding Corpus for AI-Augmented Japanese Language and Culture Learning..... 125

Abstracts of Papers

TESLA

Jaka Čibej, Simon Krek, Carole Tiberius, Irina Lobzhanidze, Verginica Barbu Mititelu, Jelena Kallas, Kertu Saul, Kadri Muischnek, Karoliina Jõgi, Ranka Stanković, Cvetana Krstev, Aleksandra Marković, Milica Ikonić Nešić, Ana Ostroški Anić, Bartłomiej Alberski, Vladimir Cvetkoski, Voula Giouli, Rusudan Makhachashvili, Olha Kanishcheva

ELEXIS-WSD 2.0: Expanding the Lexicon–Corpus Interface with New Languages and Annotation Layers 131

Milica Dinić Marinković, Cvetana Krstev, Aleksandra Marković, Ranka Stanković

Repozitorijum simila savremenog srpskog jezika – izgradnja i evaluacija mašinski potpomognute anotacije 135

Andelka Zečević, Jovana Radenović, Milica Gašić, Songbo Hu, Ranka Stanković, Mihailo Škorić

srWOZ: Towards Task-Oriented Dialogue Modelling in Serbian..... 139

Ранка Станковић, Михаило Шкорић, Рада Стujiовић

Речник савременог српског језика: од дигиталне лексикографске базе до отворене језичке инфраструктуре 142

*Јована Рађеновић, Биљана Рујевић, Александра Томашевић,
Оливера Китановић, Ранка Станковић*

**Од (слике) текста ка графу знања: дигитализација и
семантичко обogaћивање енциклопедије Свезнање.....** 144

Milena Šošić, Ranka Stanković

**Hierarchical Recognition of Moral Foundations in
Serbian Literary Corpus SrpELTeC.....** 147

Jelena Lazarević, Ranka Stanković, Olivera Kitanović

**Named Entity Recognition and Relation Extraction in
Serbian Football Text** 149

Ranka Stanković, Milena Šošić

**Lexicon-Based Exploration of Emotion Dynamics in
TEI Novels: Case Study on SrpELTeC Corpus** 151

Mihailo Škorić, Nikola Janković, Tanja Samardžić

**Comparative Intrinsic Evaluation of LLM Tokenizers
on Serbian Text** 153

Mihailo Škorić, Cosimo Palma

Wiki Dumps to Training Corpora: South Slavic Case..... 155

Language Technologies

Маријана Ђукић, Слободан Новокмет

**Поредбени фразеологизми са компонентом пас у
српском језику: корпусно заснована анализа** 158

Јелена Андоновски, Александра Тртовац

**Именице и колокације придев–именица у српском
делу паралелног корпуса часописа Инфотека:
текстометријски приступ.....** 160

Predrag Kovačević, Marko Simonović, Marija Đokić Petrović

**Hypothesis-Driven Corpus Querying at Scale: A
Method and Its Application to -nje Nominalizations in
BCMS** 162

Данка Јокић, Милица Андрић

СентИроСа: корпус за детекцију ироније и сарказма у кратким текстовима на српском језику 164

Данка Јокић, Срђан Шућур, Јелена Марковић

ADMIRE 2 Serbian: Изградња српског мултимодалног скупа идиома 166

Benedikt Perak, Dragana Špica

AI4LANG – Integrating Artificial Intelligence for the Digital Transformation of Language Teaching 168

Nevena Ceković

The ItalSerb Spoken Corpus of L2 Italian and a Paired Subcorpus of L1 Serbian..... 170

Katerina Zdravkova

Linguistic Annotation of South Slavic Languages with Large Language Models 173

Jelena Redli

Anotacija moralnih vrednosti u srpskom onlajn diskursu: izazovi i perspektive 175

Milica Jevtić, Tamara Vučenović, Miloš Stojanović, Danijela Nešić

From Literary Memory to Platformed Culture: South Slavic Diaspora Identity and Algorithmic Marginalization in Canadian Multicultural Literature 178

Andela Babić, Marko Simonović, Predrag Kovačević

Deverbal Nominalisations in the Timok Speech: A Preliminary Corpus Study 180

Jana Kuzmanova, Sonja Gievska, Katerina Zdravkova

Exploring In-Context Learning for LLM Machine Translation for Macedonian..... 182

Miloš Košprdić, Isidora Dekić, Tijana Radović

Towards Fine-Grained Annotation of Directed Hate and Abusive Speech in Serbian 184

Dragana Kunčer

Translation Direction in the Gesta regum Sclavorum: A Quantitative and Qualitative Approach	187
--	------------

Semantic Role Labeling

Svetla Koeva

Modeling Expression of Eventuality Types in Bulgarian with Large Language Models	191
---	------------

Svetlozara Leseva, Ivelina Stoyanova

Approaches to Automatic Annotation of Semantic Roles and Syntactic Relations	193
---	------------

Marina Bagi, Aleksandra Marković, Ely Matos

A FrameNet Analysis of Serbian Verbs Kidati and Cepati: Frames vs Traditional Word Sense Disambiguation	195
--	------------

Editors' Preface

The second international scientific conference, South Slavic Languages in the Digital Environment – JuDig 2026 (October 26–28, 2026, Faculty of Mining and Geology, University of Belgrade), also serves as the final conference of the TESLA project (Text Embeddings – Serbian Language Applications), supported by the Science Fund of the Republic of Serbia (Grant No. 7276). The conference programme showcases the project's research outcomes through ten papers and two workshops.

The landscape of language technologies, computational linguistics, lexicography, digital humanities, and related fields is growing rapidly. The conference serves as a platform for sharing knowledge, ideas, and technologies. This is especially important for South Slavic languages, many of which are under-resourced in language technologies and tools.

The conference received 38 submissions. After double-blind peer review and the editors' final evaluation, 36 papers were accepted for presentation at the conference, including 27 full papers and 9 extended abstracts; two papers were rejected.

JuDig 2026 brings together 93 researchers from 20 countries, representing 48 institutions. Most researchers are from Serbia (50). Croatia has 7, Slovenia 5, North Macedonia and Estonia each have four researchers, followed by Bulgaria, Brazil, and Germany, each with three. Two researchers come from Republika Srpska and Italy, while Czechia, Austria, Switzerland, the United Kingdom, Ukraine, Greece, Poland, Romania, Georgia, and the Netherlands each have one representative.

A special highlight of the conference is the keynote sessions, featuring distinguished invited speakers who will address cutting-edge topics in the field. These presentations will focus on crucial resources and infrastructures, including FrameNet (Prof. Tiago Timponi Torrent from Brazil), PressMint (Prof. Petya Osenova from Bulgaria), and the CLARIN research infrastructure (Francesca Frontini, PhD, from Italy).

The conference includes four workshops: CLASSLA-Express 3.0 as a pre-conference workshop; a workshop dedicated to FrameNet: The

FrameNet Brasil Framework; and, as mentioned, two workshops dedicated to the TESLA project: TESLA: Language Resources and TESLA: From Text to Knowledge Graphs. The conference will also feature a panel session on AI4LANG – Integrating AI for the Digital Transformation of Language Teaching.

The papers compiled in the conference proceedings present advancements in the curation, annotation, and deployment of digital language resources, with a strong emphasis on the Serbian language and the broader South Slavic linguistic context. A key thematic pillar centres on the further development of multi-layered, open-access lexical infrastructures, as demonstrated by the expansion of the ELEXIS-WSD 2.0 parallel sense-annotated corpus to support multilingual WSD, alongside the architectural transformation of the Dictionary of Contemporary Serbian Language (RSSJ) into a fully integrated digital framework connected via REST APIs to established lexical databases. Methodologically, the research showcases a clear paradigm shift toward hybrid computational strategies, combining rule-based symbolic systems and traditional corpus query platforms with state-of-the-art LLMs to automate the precise semantic annotation and evaluation of complex figurative structures, such as similes. Furthermore, the volume addresses critical resource scarcity in conversational AI by introducing srWOZ, a culturally adapted, multi-domain, task-oriented dialogue dataset specifically designed to advance token-level modelling and dialogue system evaluation in low-resource settings.

The proceedings also include papers dedicated to FrameNet, an essential resource in the NLP landscape. Developing this resource is paramount, given its critical role in advancing modern NLP systems.

The papers in the Book of Abstracts focus on the digital transformation of the humanities through the synergy of computational linguistics, NLP, and artificial intelligence. The main themes centre on creating, enhancing, and evaluating language resources, primarily specialised, parallel, and comparable corpora (covering media, onomastic, institutional, and tour-guiding discourse). The authors pay special attention to the challenges of processing under-resourced languages (such as Serbian, Croatian, and Church Slavonic), including evaluating transliteration accuracy, word sense disambiguation, and the digital preservation of specific graphic features in historical manuscripts. The authors critically assess the capabilities and stability of LLMs, specifically ChatGPT and Codex skill workflows, in complex tasks such as literary translation, lexical-semantic interpretation, and AI-assisted transcription. In conclusion, the papers strive to bridge the gap between traditional

philological analysis and modern language technologies, paving the way for open-access platforms, public databases, and advanced digital tools in both education and research.

We sincerely thank all the authors who submitted their work and contributed to the high scientific quality of the conference proceedings and this volume. We thank the members of the Programme Committee and the anonymous reviewers for their time, expertise, and dedication in conducting the double-blind peer review process.

We extend our deepest appreciation to the institutions that organised this important scientific event; besides the Faculty of Mining and Geology, these include the Language Resources and Technologies Society JeRTeH, the Faculty of Philology (University of Belgrade), the Institute for the Serbian Language of SASA, and the University Library ‘Svetozar Marković’.

We are also immensely grateful for the essential financial and institutional support provided by the Ministry of Science, Technological Development and Innovation, and the Science Fund of the Republic of Serbia. Special thanks go to the conference’s gold sponsor – The Factory Group. The Office for Information Technologies and e-Government, Envi Software Solutions, the ‘Okupljeni oko jezika’ Association, and the Institute for Standardisation of Serbia also supported JuDig 2026, and we thank them all.

Finally, we hope that conference proceedings and this volume will serve as a valuable resource for researchers, students, and practitioners, inspiring further advancements in language technologies and digital humanities.

Belgrade, October 2026
The Editors



Invited Speakers

Prof. Tiago Timponi Torrent

Professor at Universidade Federal de Juiz de Fora (UFJF)
<https://www.tiagotorrent.com/about>

Tiago Torrent is a Cognitive Linguist working on Natural Language Processing within the framework of Frame Semantics and Construction Grammar. He is head of the FrameNet Brasil Computational Linguistics Lab, PI of ReINVenTA and Professor of the Graduate Program in Linguistics at the Federal University of Juiz de Fora, Brazil. He is a Research Productivity Grantee of the Brazilian National Council for Scientific and Technological Development (CNPq) and the President of the Global FrameNet Association. He served as a guest professor in the Department of Swedish, Multilingualism and Language Technology at the University of Gothenburg.

Проф. др Тијаго Тимпони Торент

Федерални универзитет Жуиз де Фора у Бразилу (UFJF)
<https://www.tiagotorrent.com/about>

Тијаго Торент је когнитивни лингвиста који ради на обради природног језика из угла семантике оквира и конструкционе граматике. Управник је лабораторије FrameNet Brasil Computational Linguistics Lab, главни истраживач у ReINVenTA и професор на постдипломским студијама лингвистике на Федералном универзитету Жуиз де Фора у Бразилу. Добитник је гранта за продуктивност у истраживању Бразилског националног савета за научни и технолошки развој (CNPq) и председник глобалне асоцијације FrameNet. Радио је као гостујући професор на Департману за шведски језик, вишејезичност и језичке технологије на Универзитету у Гетеборгу.

Curation is also something you need: on the role of Language Resources in AI

The use of Large Language Models – recently rebranded as AI – in societally grounded scenarios such as healthcare and assistive technology has been the object of both praise and criticism. On the one hand, LLMs

create opportunities to deliver and scale healthcare and assistive technology solutions to populations which would otherwise not have access to them. On the other hand, they may introduce harmful biases, produce hallucinations and promote stereotyping and erasure of cultures.

In this talk, I claim that human-curated language resources such as FrameNet can mitigate such risks, while maintaining the scale and speed of LLM-based applications. I discuss two cases developed by the FrameNet Brasil Lab: an early warning system for Non-Communicable Diseases used in the Brazilian Universal Healthcare System and an audio description generation technology.

Кураторство је такође нешто што вам је потребно: о улози језичких ресурса у вештачкој интелигенцији

Употреба великих језичких модела (ВЈМ) – однедавно названих вештачка интелигенција (ВИ) – у друштвено утемељеним сценаријима, као што су здравствена заштита и асистивне технологије, предмет је како похвала, тако и критика. Са једне стране, ВЈМ стварају могућности за пружање решења у области здравствене заштите и асистивних технологија и њихово допирање до популације која до њих не би имала приступ. Са друге стране, они могу увести штетну пристрасност, произвести халуцинације и промовисати стереотипе и брисање читавих култура.

У овом излагању тврдим да језички ресурси које су уредили људи, као што је FrameNet, могу ублажити такве ризике, уз задржавање обима и брзине апликација заснованих на ВЈМ. Размотрићу два случаја која је развила лабораторија FrameNet Brasil Lab: систем раног упозоравања на немикробиолошке (незаразне) болести који се користи у бразилском универзалном здравственом систему и технологију за генерисање аудио дескрипције.

Prof. Petya Osenova

Faculty of Slavic Studies at Sofia University “St. Kl. Ohridski” and IICT-BAS

petya@bultreebank.org

Dr Petya Osenova is a professor in Contemporary Bulgarian Grammar (morphology, syntax and corpus linguistics) in the Faculty of Slavic Studies at Sofia University “St. Kl. Ohridski” and a senior researcher in the area of Language Technologies in the Department of Linguistic Modelling and Knowledge Processing at the Institute of Information and Communication Technologies, Bulgarian Academy of Sciences. Her scientific interests include formal and computational linguistics, language resources, natural language processing, and the grammar-lexicon interface.

Petya Osenova was a key person in several EU projects related to eLearning, Machine Translation, and Language resources (EuroMatrixPlus, AsIsKnown, LTfLL, QTLeap, EUCases, among others). She is the responsible person for the language resources in CLaDA-BG – the CLARIN and DARIAH joint framework in Bulgaria. Petya Osenova is the Bulgarian representative at the CLARIN User Involvement Committee and its chair from 2026 in CLARIN-ERIC. Petya Osenova was a co-coordinator of the ParlaMint project (the first CLARIN flagship project 2020-2023) and is a participant in the follow-up OSCARS project ParlaCAP (2005 - 2007). She is the coordinator of the current CLARIN Flagship project PressMint (2026 - 2027).

Petya Osenova specialised in computational linguistics as a postdoctoral fellow at the University of Tübingen, Germany (2003) and at the University of Groningen, the Netherlands (2004); as a Fulbright scholar at Stanford University, the USA (2010). Her postdoctoral projects were related to the construction of the BulTreeBank, the quantitative measurement of Bulgarian dialects and the Resource grammar construction for Bulgarian.

In 2018, Petya Osenova received the award from Clarivate Analytics for excellence in scientific research in South-Eastern Europe in the area of Social Sciences.

Проф. др Петја Осенова

Факултет за словенску филологију, Универзитет у Софији „Св. Климент Охридски“ и Институт за информационо-комуникационе технологије Бугарске академије наука (ИСТ-BAS)

petya@bultreebank.org

Др Петја Осенова (Petya Osenova) редовни је професор Савремене бугарске граматике (морфологија, синтакса и корпусна лингвистика) на Факултету за словенску филологију Универзитета у Софији „Св. Климент Охридски“ и виши научни сарадник у области језичких технологија на Одсеку за лингвистичко моделирање и обраду знања при Институту за информационо-комуникационе технологије, Бугарске академије наука (ИСТ-BAS). Њени научни интереси обухватају области формалне и рачунарске лингвистике, језичких ресурса, обраде природног језика, као и однос граматике и лексикона.

Петја Осенова је била кључни учесник у низу пројеката Европске уније везаних за електронско учење (е-учење), машинско превођење и језичке ресурсе (између осталих: EuroMatrixPlus, AsIsKnown, LTfLL, QTLear и EUCases). Главна је одговорна особа за језичке ресурсе у оквиру CLaDA-BG – заједничке мреже инфраструктура CLARIN и DARIAH у Бугарској. Представник је Бугарске у CLARIN Комитету за укључивање корисника (CLARIN User Involvement Committee), а од 2026. године обавља и функцију председавајуће овог тела у склопу конзорцијума CLARIN-ERIC. Петја Осенова је била координатор пројекта ParlaMint (првог водећег пројекта мреже CLARIN у периоду 2020–2023) и учествује у његовом наставку – пројекту ParlaCAP у оквиру иницијативе OSCARS (2025–2027). Такође је координатор актуелног водећег CLARIN пројекта PressMint (2026–2027).

На пољу рачунарске лингвистике усавршавала се као постдокторанд на Универзитету у Тибингену у Немачкој (2003. године), на Универзитету у Гронингену у Холандији (2004. године), као и на Универзитету Станфорд у САД (2010. године) као добитник Фулбрајтове стипендије. Њени постдокторски пројекти били су усмерени на изградњу корпуса BulTreeBank, квантитативно мерење бугарских дијалеката и креирање граматике језичких ресурса за бугарски језик.

Године 2018. Петја Осенова је добила награду издавачке куће Clarivate Analytics за изузетност у научном истраживању у Југоисточној Европи у области друштвених наука.

PressMint: Towards Interoperable Corpora of Historical Newspapers

The talk will focus on PressMint – an ongoing CLARIN ERIC flagship project (2025-2027).

PressMint aims to compile a multilingual, comparable, annotated, translated and interoperable set of corpora of European historical newspapers, centred around the start of the 20th century. The participating parties feature 21 institutions across 18 countries.

I will discuss its infrastructure for historical newspaper creation, which includes: a source and license management protocol based on the participating teams' experience; encoding guidelines; a linguistic annotation schema; validation, conversion, extraction, and enrichment scripts; etc.

The challenges and possible solutions behind the provision of historical newspaper corpora will be outlined as well.

PressMint: ка интероперабилним корпусима историјских новина

Предавање ће бити фокусирано на PressMint – актуелни водећи пројекат (flagship project) инфраструктуре CLARIN ERIC (2025–2027).

PressMint има за циљ да прикупи вишејезични, упоредив, аотиран, преведен и интероперабилан скуп корпуса европских историјских новина с почетка 20. века. Међу учесницима је 21 институција из 18 земаља.

Говорићу о инфраструктури за креирање корпуса историјских новина која укључује: протокол за управљање изворима и лиценцама, заснован на искуству тимова који учествују; смернице за кодирање; схему лингвистичке аотације; скрипте за валидацију, конверзију, екстракцију и обогаћивање података, итд.

Такође ће бити представљени изазови и могућа решења која прате прављење корпуса историјских новина.

Dr Francesca Frontini

Researcher at the Istituto di Linguistica Computazionale "A. Zampolli" of the CNR in Pisa

<https://sites.google.com/site/francescafrontini/>

Francesca Frontini is a researcher at the Institute for Computational Linguistics “Antonio Zampolli” (ILC-CNR) in Pisa, Italy. Her research focuses on language resources, corpus linguistics, and natural language processing, with particular attention to the creation, annotation, and interoperability of linguistic datasets. She was a member of the Board of Directors of the CLARIN research infrastructure from 2021 to 2025. She is currently involved in the European initiative LLMs4EU, where she contributes to activities on language resources and AI.

Др Франческа Фронтини

Истраживач у Институту за рачунарску лингвистику „А. Замполи“ CNR-а у Пизи

<https://sites.google.com/site/francescafrontini/>

Франческа Фронтини је истраживач на Институту за компјутерску лингвистику „Antonio Zampolli” (ILC-CNR) у Пизи, Италија. Њено истраживање је усмерено на језичке ресурсе, корпусну лингвистику и обраду природног језика, а посебно креирању, анотацији и интероперабилности лингвистичких скупова података. Била је члан Управног одбора истраживачке инфраструктуре CLARIN од 2021. до 2025. године. Тренутно је ангажована на европској иницијативи LLMs4EU, где доприноси активностима у области језичких ресурса и вештачке интелигенције (ВИ).

CLARIN and the Emerging European Ecosystem for Language Data and Technologies: The Italian Perspective

Language resources and technologies are essential for research, innovation, and cultural heritage in the digital age. European research infrastructures such as CLARIN (Common Language Resources and Technology Infrastructure) support the creation, sharing, and long-term preservation of language data, promoting interoperability and the FAIR principles.

In this talk, I will introduce the CLARIN infrastructure and discuss the benefits of participating in such initiatives for national research communities. Drawing on Italy's experience and participation in several national and European projects, I will present CLARIN in the broader context of the evolving European ecosystem for research data and AI, including the European Open Science Cloud (EOSC) and emerging initiatives such as the Language Data Space, ALT-EDIC, and the European Collaborative Cloud for Cultural Heritage.

CLARIN и европски екосистем за језичке податке и технологије у успону: италијанска перспектива

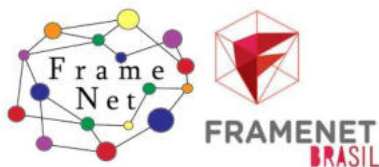
Језички ресурси и технологије кључни су за истраживање, иновације и културну баштину у дигиталном добу. Европске истраживачке инфраструктуре, као што је CLARIN (Заједничка инфраструктура за језичке ресурсе и технологију), подржавају креирање, дељење и дугорочно очување језичких података, промовишући интероперабилност и FAIR принципе.

У овом излагању представићу инфраструктуру CLARIN и говорићу о предностима које учешће у таквим иницијативама може донети националним истраживачким заједницама. Ослањајући се на искуство Италије и учешће у неколико националних и европских пројеката, приказаћу CLARIN у ширем контексту еволуирајућег европског екосистема за истраживачке податке и вештачку интелигенцију (AI). То укључује Европски облак за отворену науку (EOSC) и нове иницијативе као што су Простор за језичке податке (Language Data Space), ALT-EDIC и Европски колаборативни облак за културну баштину.



Workshops

Building Enriched Frame-Based Resources: The FrameNet Brasil Framework



26. 10. 2026, 15.00–18.00 (Room 151)

FrameNet is the computational implementation of Frame Semantics theory. According to Fillmore (1985), as one of the Semantics of Understanding, such a theory aims “to uncover the nature of the relationship between linguistic texts and the interpreter's full understanding of texts in their contexts” (Fillmore, 1985, p.231). However, the original Berkeley FrameNet model falls short in what concerns the representation of several types of contextual information which are key for building frame-based semantic models (Torrent et al., 2022). In this context, FrameNet Brasil (FNBr) has been extending the Berkeley FrameNet model not only to include another language—Brazilian Portuguese—but also to represent additional contextual information which can be derived from additional relations in the database and new types of multimodal annotation.

The operationalisation of FNBr is carried out through WebTool. The WebTool is a web application developed "in-house" that allows remote access to the FNBr database. It is divided into 3 areas:

- reports and graphs on the data (Frames, LUs, Constructions);
- lexical and multimodal annotation (images and videos); and
- maintenance of the system's internal tables.

Two aspects of the WebTool's design are foundational to enriching the FrameNet model:

- the implementation of an entity-based data model and
- support for multilinguality.

In the WebTool, all elements (Frames, FEs, LUs, SemanticTypes, etc.) are treated as Entities that can relate to one another. This allows flexibility in creating new relations without the need for structural changes to the database. As a multilingual tool, the WebTool allows the names of the elements to be specific to each language, enabling the tool to be used by groups from various countries. In this workshop, we provide an overview of the functionalities of the WebTool for FrameNet creation.

In this workshop, an overview will be provided of the functionalities of the WebTool for FrameNet creation, focusing on its entity-based architecture, multilingual support, and facilities for lexical and multimodal annotation.

Participants will gain hands-on experience in navigating the FrameNet Brasil infrastructure, exploring frame and lexical unit repositories, creating and managing frame-semantic resources, and annotating textual, visual, and audiovisual data. The workshop will also demonstrate how WebTool supports the development of multilingual FrameNets and frame-based semantic models.

By the end of the workshop, participants will be able to:

- navigate and query FrameNet resources;
- create and edit frames, frame elements, and lexical units;
- perform lexical and multimodal annotation;
- explore multilingual FrameNet development;
- understand how frame-based semantic resources can support NLP applications and linguistic research.

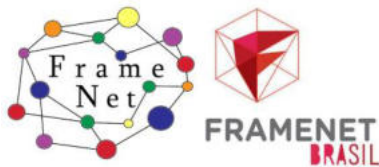
Who is it for?

The workshops are intended for students, linguists, lexicographers, digital humanities researchers, and anyone interested in frame semantics, FrameNet, and WebTool.

Organizers:

- Tiago Timponi Torrent, UFJF | Linguística - FrameNet Brasil, tiagotorrent.com
- Ely Matos, Universidade Federal de Juiz de Fora, <https://ematos.net/en/home-2/>
- Marcelo Viridiano, National Institute for Science and Technology on Responsible AI for Computational Linguistics and Information Treatment and Dissemination

Изградња обogaћених ресурса заснованих на оквирима: окружење FrameNet Brasil



26. 10. 2026, 15.00–18.00 (Соба 151)

FrameNet представља рачунарску имплементацију теорије семантике фрејмова. Према Филмору (Fillmore, 1985), као једна од теорија семантике разумевања, ова теорија има за циљ „да открије природу односа између језичких текстова и потпуног разумевања текстова у њиховом контексту од стране тумача“ (Fillmore, 1985, стр. 231). Међутим, изворни модел Berkeley FrameNet-а не омогућава адекватно представљање више врста контекстуалних информација које су кључне за изградњу семантичких модела заснованих на фрејмовима (Torrent et al., 2022). У том контексту, FrameNet Brasil (FNBr) проширује Berkeley FrameNet модел не само укључивањем још једног језика — бразилског португалског — већ и представљањем додатних контекстуалних информација које се могу извести из нових типова односа у бази података и нових облика мултимодалне анотације.

Операционализација FNBr-а остварује се путем алата WebTool. WebTool је веб апликација развијена у оквиру пројекта која омогућава удаљени приступ FNBr бази података. Подељена је на три целине:

- извештаје и графичке приказе података (фрејмови, лексичке јединице, конструкције);
- лексичку и мултимодалну анотацију (слике и видео-записи);
- одржавање интерних табела система.

Два аспекта дизајна WebTool-а од суштинског су значаја за проширење FrameNet модела:

- примена модела података заснованог на ентитетима;
- подршка вишејезичности.

У оквиру WebTool-а сви елементи (фрејмови, елементи фрејма, лексичке јединице, семантички типови и др.) третирају се као ентитети који могу бити међусобно повезани. Овакав приступ омогућава флексибилно увођење нових односа без потребе за структурним изменама базе података. Као вишејезички алат, WebTool омогућава да називи елемената буду прилагођени појединачним

језицима, што омогућава његову употребу у различитим националним и истраживачким окружењима.

На радионици ће бити представљене могућности WebTool-а за изградњу и развој FrameNet ресурса, са посебним освртом на његову архитектуру засновану на ентитетима, подршку вишејезичности и функционалности за лексичку и мултимодалну анотацију.

Учесници ће кроз практичан рад стећи искуство у коришћењу инфраструктуре FrameNet Brasil, истраживању репозиторијума фрејмова и лексичких јединица, креирању и управљању фрејмско-семантичким ресурсима, као и анотацији текстуалних, визуелних и аудио-визуелних података. Радионица ће такође показати како WebTool подржава развој вишејезичних FrameNet ресурса и семантичких модела заснованих на фрејмовима.

По завршетку радионице учесници ће бити оспособљени да:

- претражују и користе FrameNet ресурсе;
- креирају и уређују фрејмове, елементе фрејма и лексичке јединице;
- спроводе лексичку и мултимодалну анотацију;
- истражују развој вишејезичних FrameNet ресурса;
- разумеју на који начин фрејмско-семантички ресурси могу подржати примене у обради природног језика и лингвистичка истраживања.

Коме је радионица намењена?

Радионица је намењена студентима, лингвистима, лексикографима, истраживачима у области дигиталне хуманистике и свима који су заинтересовани за семантику фрејмова, FrameNet и WebTool.

Организатори:

- Tiago Timponi Torrent, UFJF | Linguística - FrameNet Brasil, tiagotorrent.com
- Ely Matos, Universidade Federal de Juiz de Fora, <https://ematos.net/en/home-2/>
- Marcelo Viridiano, National Institute for Science and Technology on Responsible AI for Computational Linguistics and Information Treatment and Dissemination

Spoken and web corpora in the analysis of language variation



CLASSLA-express 3.0

workshops on using CLARIN.SI resources in language research

Pre-conference workshop 25. 10. 2026, 10.00–14.00 (Room 151)

We are pleased to announce that the Croatian Society for Applied Linguistics is once again collaborating with [CLARIN.SI](https://clarin.si) to organise the CLASSLA-Express workshop series.

Following the successful editions of [CLASSLA-Express 1.0](#) and [2.0](#), CLASSLA-Express 3.0 introduces a new focus on spoken and web corpora and their role in the analysis of language variation.

Workshop content

The workshops offer hands-on work with spoken and web corpora. In the first part, participants will explore the [ParlaSpeech](#) parliamentary speech corpus using the [NoSketchEngine concordancer](#), analyse concordances, and extract variables and speech features with the help of Praat.

To get a first look at the ParlaSpeech corpus, you can watch the following short video guides:

[Introducing ParlaSpeech](#)
[Exploring ParlaSpeech](#)

In the second part, participants will work with [CLASSLA-web](#) corpora, build subcorpora using the [NoSketchEngineLog concordancer](#), and compare lexical patterns across different genres and domains.

To get familiar with CLASSLA-web corpora in advance, we recommend the following video guide:

[Introducing CLASSLA-web](#)

Participants will learn how to:

- formulate corpus queries
- analyse concordances and metadata
- extract and interpret linguistic variables from concordance data
- build subcorpora
- compare lexical patterns
- explore features of spoken language

Who is it for?

The workshops are intended for students, linguists, phoneticians, lexicographers, language teachers, digital humanities researchers, and anyone interested in corpus-based language research.

Workshops schedule 2026

October – Belgrade, Serbia (pre-conference workshop, JuDig)

Organizers:

- Ivana Filipović Petrović, Croatian Academy of Sciences and Arts, Croatia
- Nikola Ljubešić, Jožef Stefan Institute, University of Ljubljana, Institute of Contemporary History, Slovenia
- Ivan Porupski, Jožef Stefan Institute, Slovenia
- Isabell Furkert, University of Ljubljana, Slovenia
- Tanja Samardžić, IDSIA, USI-SUPSI, Lugano, Switzerland
- Kristina Ilić, Metropolitan University, Serbia
- Robert Sviben, Institute of Croatian Language, Croatia

Workshop Program

- 10.00–10.15  **Opening of the Workshop and Introduction to the CLARIN.SI Research Infrastructure**
 Tanja Samardžić
- 10.15–11.30  **ParlaSpeech: Searching and Analysing Parliamentary Speech Corpora**
 Tanja Samardžić
- 11.30–12.00   Coffee Break
- 12.00–13.00  **CLASSLA-web: Building Subcorpora and Comparing Genres**
 Kristina Ilić
- 13.00–14.00  **Phraseological Constructions in Corpora: CQL Queries and Pattern Analysis**
 Kristina Ilić

Говорни и веб-корпуси у истраживању језичке варијације



CLASSLA-Express 3.0

радионице о употреби ресурса CLARIN.SI у језичким истраживањима

Преконференцијска радионица, 25. 10. 2026, 10.00–14.00 (Соба 151)

Са задовољством најављујемо да Хрватско друштво за примењену лингвистику и ове године учествује у организацији радионица CLASSLA-Express, у сарадњи са инфраструктуром CLARIN.SI

Након успешних издања CLASSLA-Express 1.0 и 2.0., CLASSLA-Express 3.0 доноси нови угао гледања на говорне и веб корпусе и њихову примену у анализи језичке варијације.

Садржај радионица

Радионице нуде практичан рад с говорним и веб корпусима. У првом делу учесници ће претраживати парламентарни говорни корпус ParlaSpeech помоћу алата NoSketchEngine, анализирати конкорданце и издвајати варијабле и говорна обележја уз помоћ алата Praat.

За упознавање са корпусом ParlaSpeech можете унапред погледати кратка видео-упутства:

[Увод у ParlaSpeech](#)

[Рад са ParlaSpeech-ом](#)

У другом делу учесници ће радити са корпусима CLASSLA-web, израђивати поткорпусе у алату NoSketch Engine и упоређивати лексичке обрасце између различитих жанрова и домена.

За упознавање са корпусима CLASSLA-web препоручујемо следеће видео-упутство:

[Увод у CLASSLA-веб](#)

Током радионица учесници ће научити како да:

- постављају корпусне упите;
- анализирају конкорданце и метаподатке;
- издвајају и тумаче језичке варијабле из конкорданци;
- израђују поткорпусе;
- упоређују лексичке обрасце;
- истражују обележја говорног језика.

Коме су радионице намењене?

- Радионице су намењене студентима, лингвистима, фонетичарима, лексикографима, наставницима језика, истраживачима у области дигиталне хуманистике, као и свима који су заинтересовани за корпусно заснована језичка истраживања.

Организатори:

- Ивана Филиповић Петровић, Хрватска академија знаности и умјетности, Хрватска
- Никола Љубешић, Институт „Јожеф Стефан“, Универзитет у Љубљани, Институт за савремену историју, Словенија
- Иван Порупски, Институт „Јожеф Стефан“, Словенија
- Изабел Фуркерт, Универзитет у Љубљани, Словенија
- Тања Самарцић, IDSIA, USI-SUPSI, Швајцарска
- Кристина Илић, Универзитет Метрополитан Београд, Србија
- Роберт Свибен, Институт за хрватски језик, Хрватска

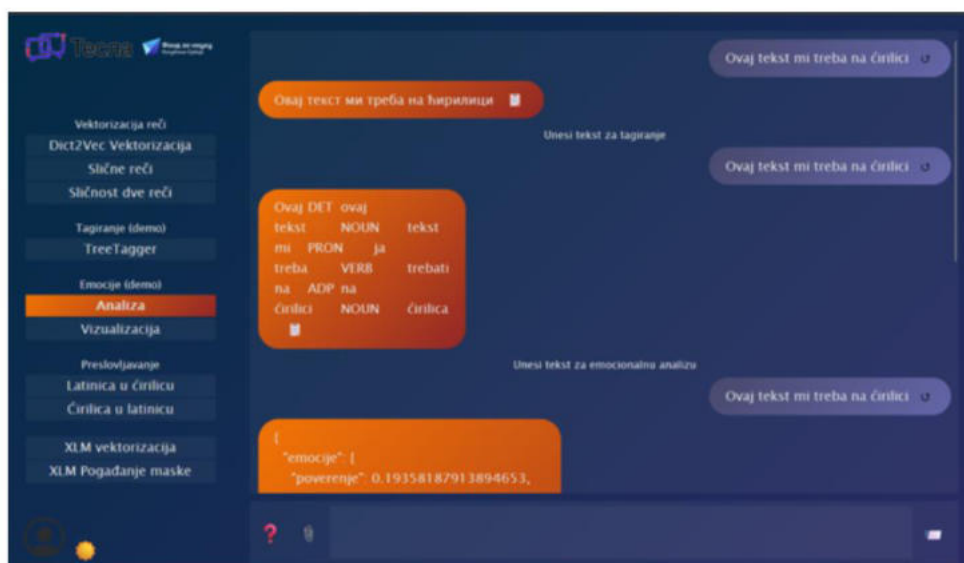
Програм радионице

- 10.00–10.15  **Отварање радионице и презентација истраживачке инфраструктуре CLARIN.SI**
 Тања Самарџић
- 10.15–11.30  **ParlaSpeech: претраживање и анализа говорног корпуса**
 Тања Самарџић
- 11.30–12.00   Пауза за кафу
- 12.00–13.00  **CLASSLA-web: израда поткорпуса и упоређивање жанрова**
 Кристина Илић
- 13.00–14.00  **Фраземске конструкције у корпусу: претраживање и анализа образаца**
 Кристина Илић

TESLA: Language Resources. Use of language resources developed within the framework of the TESLA project with a focus on the TESLA portal



28. October, 2026, 10.30–13.30 (Room 151)



Workshop Description

A number of **language resources and tools for text processing in Serbian** were developed within the **TESLA** project, with the aim of supporting practical application in a digital environment. This workshop aims to provide participants with a comprehensive insight into the capabilities of some of the tools, as well as to show how they can be used through an **open portal** intended for the wider community.

Special emphasis is placed on **modern text processing**, which includes normalisation, transliteration, annotation, vector analysis and the

application of language models. Participants will have the opportunity to familiarise themselves with how these tools enable this process.

The workshop is designed to offer participants a combination of theoretical introduction and **practical work**, allowing them to acquire concrete knowledge and skills through interactive examples. In this way, participants will not only understand how the tools work, but will also be able to apply them themselves.

Practical Session

Participants will test individual portal functionalities such as text normalisation and transliteration, annotation, vector analysis, and the use of language models for text processing.

They will:

- transliterate text from Cyrillic to Latin and vice versa with a focus on special examples and exceptions;
- annotate text in the form of part-of-speech tagging and lemmatisation;
- calculate word similarity in Serbian using various algorithms for static vectorisation;
- convert text to vector and perform vector operations;
- vectorise text in separate vector spaces (emotion and evocation of emotions);
- visualise and analyse the obtained results.

Who is the workshop for?

The workshop is intended for everyone who is interested in language resources and text processing in the Serbian language.

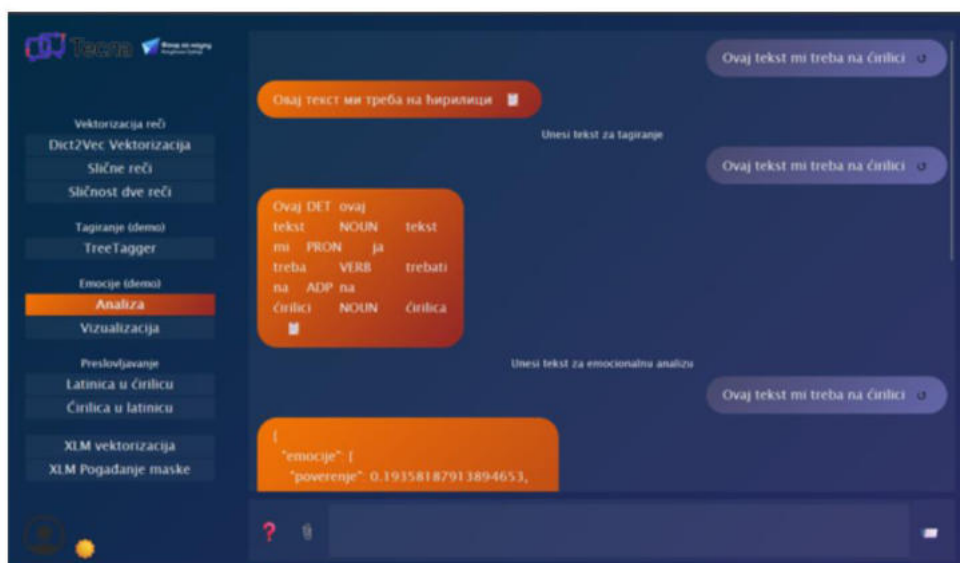
Organizers:

- Mihailo Škorić, PhD, University of Belgrade – Faculty of Mining and Geology
- Biljana Rujević, PhD, University of Belgrade – Faculty of Mining and Geology
- Milica Dinić Marinković, PhD, University of Belgrade – Faculty of Philology
- Prof. Ranka Stanković, University of Belgrade – Faculty of Mining and Geology
- MSc Saša Petalinkar, University of Belgrade – Faculty of Mining and Geology

ТЕСЛА: Језички ресурси. Употреба језичких ресурса развијених у оквиру пројекта ТЕСЛА са фокусом на ТЕСЛА портал



28. октобар, 2026, 10.30–13.30 (Соба 151)



Опис радионице

У оквиру пројекта **ТЕСЛА** развијени су бројни **језички ресурси и алати за обраду текста на српском језику**, са циљем да се подржи практична примена у дигиталном окружењу. Циљ радионице је да учесницима пружи свеобухватан увид у могућности ових алата и покаже како се они могу користити путем **отвореног портала** намењеног широј заједници.

Посебан акценат ставља се на **савремене приступе обради текста**, који обухватају нормализацију, транслитерацију, анотацију, векторску анализу и примену језичких модела.

Конципирана као спој кратког теоријског увода и **практичног рада**, радионица ће учесницима омогућити да кроз интерактивне примере стекну конкретна знања и вештине у коришћењу ТЕСЛА ресурса и алата, и оспособе се за њихову самосталну примену у различитим задацима обраде текста.

Практичан део

Полазници ће тестирати појединачне функционалности портала попут нормализације и транслитерације текста, анотације, векторске анализе и употребе језичких модела у обради текста

Учесници ће:

- пресловљавати текст из ћирилице у латиницу и обнурто са фокусом на посебне примере и изузетке;
- анотирати текст - обележавање врста речи и лематизација;
- израчунавати сличност речи на српском језику користећи различите алгоритме за статичку векторизацију;
- претварати текст у вектор и вршити операције над векторима;
- векторизовати текст у специјалним векторским просторима (емоција и евокација емоција);
- визуализовати и анализирати добијене резултате.

Коме је радионица намењена?

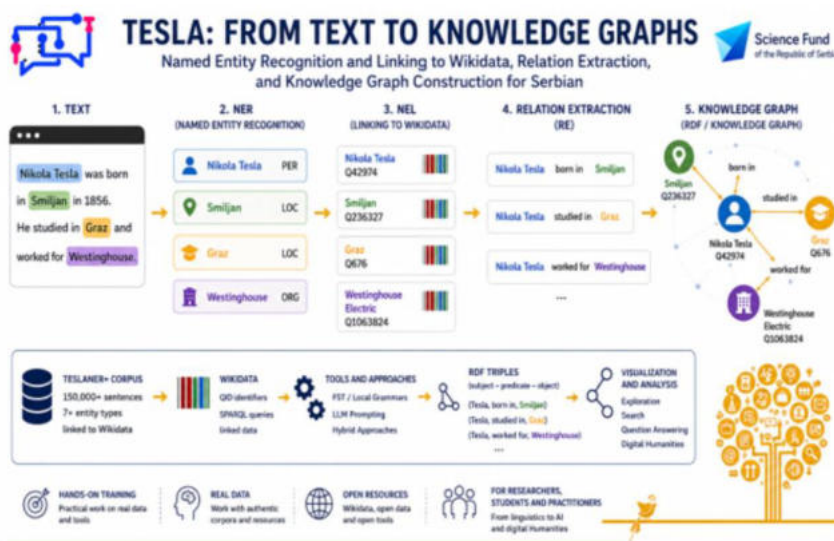
Радионица је намењена свима који су заинтересовани за језичке ресурсе и обраду текста на српском језику.

Организатори:

- др Михаило Шкорић, Универзитет у Београду – Рударско-геолошки факултет
- др Биљана Рујевић, Универзитет у Београду – Рударско-геолошки факултет
- доц. др Милица Динић Маринковић, Универзитет у Београду – Филолошки факултет
- проф. др Ранка Станковић, Универзитет у Београду – Рударско-геолошки факултет
- мрр Саша Петалинкар, Универзитет у Београду – Рударско-геолошки факултет

TESLA: From Text to Knowledge Graphs

Named Entity Recognition and Linking to Wikidata, Relation Extraction, and Knowledge Graph Construction for Serbian



27. 10. 2026, 10.30–13.30 (Room 151)

Workshop Description

Within the TESLA project, advanced resources and tools have been developed for Named Entity Recognition (NER), Named Entity Linking (NEL), and Semantic Relation Extraction (RE). Although these tasks are often studied independently, their integration enables the automatic construction of knowledge graphs from unstructured texts.

This workshop presents a complete workflow for transforming text into structured knowledge using language resources for Serbian. Participants will be introduced to the TeslaNER+ corpus, which contains more than 150,000 sentences annotated with named entities linked to Wikidata, as well as methods for relation extraction based on local grammars, finite-state transducers (FSTs), and large language models (LLMs).

Special attention will be devoted to linking entities to Wikidata, working with QID identifiers, retrieving additional information through SPARQL queries, and enriching entity descriptions using linked open data resources.

In the final part of the workshop, participants will learn how to automatically construct knowledge graphs from extracted entities and relations, represent them as RDF triples, and visualise them using graph exploration and analysis tools.

Learning Outcomes

By the end of the workshop, participants will be able to:

- recognise named entities in text;
- link entities to Wikidata (Named Entity Linking);
- use Wikidata QIDs and SPARQL queries;
- extract semantic relations between entities;
- work with relation extraction models;
- generate RDF triples in the form subject–predicate–object;
- construct and enrich knowledge graphs;
- visualise and analyse knowledge graphs;
- apply knowledge graphs in information retrieval, question answering, and digital humanities.

Hands-on Session

Participants will work with real-world examples from different domains, including literature, biographies, history, and news, using authentic data from Serbian language corpora.

The complete processing pipeline will be demonstrated:

Text → NER → NEL → RE → RDF Triples → Knowledge Graph

Target Audience

The workshop is intended for students, linguists, lexicographers, researchers in natural language processing, digital humanities scholars, librarians, information scientists, and anyone interested in semantic technologies, linked open data, and knowledge graphs.

Prerequisites

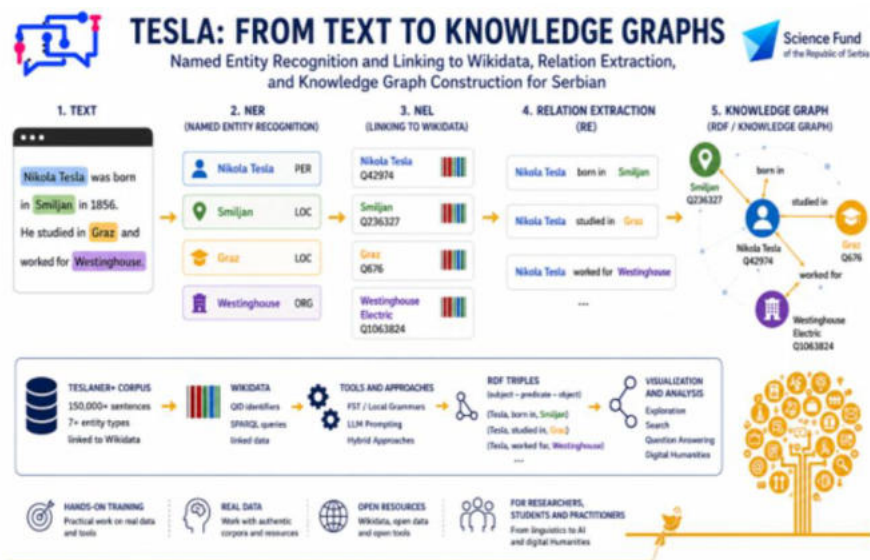
Basic familiarity with text data and corpora is recommended. No prior experience with Wikidata, RDF, or knowledge graphs is required.

Organizers

- Prof. Ranka Stanković, University of Belgrade – Faculty of Mining and Geology
- MSc Milica Ikonić Nešić, University of Belgrade – Faculty of Philology
- MSc Jovana Rađenović, University of Belgrade – Faculty of Mining and Geology
- Dr Olivera Kitanović, University of Belgrade – Faculty of Mining and Geology

ТЕСЛА: Од текста до графова знања

препознавање и повезивање именованих ентитета са Википодацима
екстракција релација и изградња графова знања за српски језик



27. 10. 2026, 10.30–13.30 (Соба 151)

Опис радионице

У оквиру пројекта ТЕСЛА развијени су напредни ресурси и алати за препознавање именованих ентитета (NER), њихово повезивање са базама знања (NEL) и екстракцију семантичких релација (RE). Иако су ови задаци често проучавани одвојено, њихова интеграција омогућава аутоматску изградњу графова знања из неструктурираних текстова.

Ова радионица пружа увид у комплетан процес трансформације текста у структурирано знање, заснован на ресурсима за српски језик. Полазници ће се упознати са корпусом TeslaNER+, који садржи више од 150.000 реченица са именованим ентитетима повезаним са Vikipodacima (Wikidata), као и са методама за екстракцију релација применом локалних граматика, коначних аутомата (FST) и великих језичких модела (LLM).

Посебна пажња биће посвећена повезивању ентитета са Википодацима, коришћењу QID идентификатора, преузимању

додатних информација путем SPARQL упита и проширивању знања о ентитетима кроз повезане ресурсе отворених података.

У завршном делу радионице учесници ће научити како да аутоматски конструишу графове знања из издвојених ентитета и релација, како да их представе у RDF формату и визуализују помоћу алата за анализу и истраживање графова.

Током радионице учесници ће научити како да:

- препознају именоване ентитете у тексту;
- повезују ентитете са Википодацима (Named Entity Linking);
- користе Wikidata QID идентификаторе и SPARQL упите;
- екстрахују семантичке релације између ентитета;
- приступе моделима за екстракцију релација;
- формирају RDF тројке облика субјект–релација–објект;
- конструишу и обогаћују графове знања;
- визуализују и анализирају графове знања;
- користе графове знања у претрази информација, питањима и одговорима и дигиталној хуманистици.

Практичан део

Полазници ће радити на примерима из различитих домена (књижевност, биографије, историја, вести), користећи стварне податке из корпуса.

Биће приказан комплетан ток обраде:

Текст → NER → NEL → RE → RDF тројке → Граф знања

Коме је радионица намењена?

Радионица је намењена студентима, лингвистима, лексикографима, истраживачима у области обраде природног језика, дигиталне хуманистике, библиотекарства и информационих наука, као и свима заинтересованима за семантичке технологије, повезане отворене податке и графове знања.

Предуслови

Основно познавање рада са текстуалним подацима и корпуса. Није неопходно претходно искуство са Википодацима, RDF-ом или графовима знања.

Организатори:

- Проф. др Ранка Станковић, Универзитет у Београду - Рударско-геолошки факултет
- мср Милица Иконић Нешић, Универзитет у Београду - Филолошки факултет
- мср Јована Рађеновић, Универзитет у Београду - Рударско-геолошки факултет
- др Оливера Китановић, Универзитет у Београду - Рударско-геолошки факултет

AI4LANG – Integrating Artificial Intelligence for the Digital Transformation of Language Teaching

The rapid advancement of generative AI and natural language processing has created transformative opportunities for language education, yet the integration of these technologies into pedagogically coherent, CEFR-aligned learning environments remains a significant challenge. This panel presents AI4LANG, a multinational research and development initiative that expands a production-ready AI Language Tutor platform across three typologically distinct languages – English, Japanese, and Croatian – addressing the global need for adaptive, AI-mediated language instruction that serves both high-resource and underrepresented languages. The platform’s architecture is built on a multi-agent orchestration paradigm where specialised AI agents handle distinct pedagogical functions – exercise generation, pronunciation evaluation, grammar feedback, and learning path adaptation – coordinated through a central orchestrator operating over a dual-database knowledge layer combining vector embeddings with structured lexical and grammatical graphs. We present the technical architecture, the CEFR-aligned pedagogical framework underpinning adaptive exercise generation, and the methodology for creating language-specific knowledge resources through LLM-assisted inference validated by expert review.

AI4LANG – Интеграција умјетне интелигенције за дигиталну трансформацију наставе језика

Брз напредак генеративне умјетне интелигенције и обраде природног језика отворио је могућности за корјенито унапређење наставе језика, но интеграција тих технологија у педагошки усклађена окружења за учење, утемељена на Заједничком еуропском референтном оквиру за језике (ЦЕФР), и даље представља значајан изазов. На овом панелу представљамо AI4LANG, међународну истраживачко-развојну иницијативу која проширује платформу AI Language Tutor, спремну за продукцијску употребу, на три типолошки различита језика – енглески, јапански и хрватски – одговарајући на глобалну потребу за прилагодљивом наставом језика посредованом умјетном интелигенцијом, намијењеном и језицима с богатим ресурсима и недовољно заступљеним језицима.

Архитектура платформе темељи се на парадигми вишеагентне оркестрације, у којој специјализирани агенти умјетне интелигенције обављају различите педагошке функције – генерирање вјежби, процјену изговора, пружање повратних информација о граматички и прилагодбу путање учења. Њихов рад координира средишњи оркестратор који дјелује над слојем знања утемељеним на двјема базама података, комбинирајући векторске репрезентације са структурираним лексичким и граматичким графовима.

Представљамо техничку архитектуру, педагошки оквир усклађен са ЦЕФР-ом на којем се темељи прилагодљиво генерирање вјежби, те методологију израде језично специфичних ресурса знања закључивањем уз помоћ великих језичких модела (LLM), чији се резултати потврђују стручном провјером. Допринос панела јест модел за развој вишејезичних суштава за поучавање утемељених на умјетној интелигенцији, који се може примијенити и у другим срединама, а који у различитим језичним контекстима уравнотежује технолошку напредност и педагошку утемељеност.

Momo Kapor u digitalnom okruženju: Književno prevođenje i ChatGPT

Olja Perišić¹[\[0000-0002-1219-862X\]](#)

¹ Univerzitet u Torinu, Departman za strane jezike, književnost i moderne kulture,
Italija, olja.perisic@unito.it

Apstrakt

Tema rada je ispitivanje mogućnosti i ograničenja modela ChatGPT u prevođenju književnih tekstova sa srpskog na italijanski, na primeru uvodnog citata romana *Beleške jedne Ane Mome Katora*. Analiza će biti usmerena na one elemente koji predstavljaju poseban izazov za generativno prevođenje pomoću velikih jezičkih modela, posebno u slučaju jezika s manjim brojem digitalnih resursa, kod kojih se češće uočavaju problemi vezani za verodostojnost ponuđenih rešenja i njihovu korektnu interpretaciju u odnosu na kontekst (Hendy et al., 2023; Guerreiro et al., 2023). U fokusu analize biće sleng u dijahroniji, idiomatski izrazi, realije beogradskog prostora sedamdesetih godina i pojedina fonetska obeležja teksta. U tu svrhu formulisana su četiri tipa promptova *basic*, *transbrief*, *author* i *translator* (He, 2024) koji su najpre testirani na modelu ChatGPT 5.4, dok su dodatne provere, zasnovane na ponavljanju istih tipova promptova od strane različitih korisnika sprovedene pomoću modela ChatGPT 5.5 kako bi se ispitala stabilnost dobijenih prevodilačkih rešenja. Rezultati pokazuju da ChatGPT generiše prividno razumljive i čitljive tekstove koji često sadrže kritične elemente na semantičkom, pragmatičnom i kulturnom planu. Ponovljeno promptovanje dovelo je do veće leksičke raznovrsnosti i poboljšanja pojedinačnih rešenja koja međutim nisu sistematična, niti stabilna. Zaključak je da obrada kulturno, dijahronijski i stilski obeleženih elemenata u književnim tekstovima zahteva ne samo postojanje specijalizovanih digitalnih resursa za srpski jezik, već i njihovo promišljeno uključivanje u postupke integrisanja, prilagođavanja i evaluacije velikih jezičkih modela.

Ključne reči: ChatGPT, književno prevođenje, Momo Kapor

1. Uvod

U radu se ispituju mogućnosti i ograničenja modela ChatGPT u prevođenju književnih tekstova sa srpskog na italijanski. Kao primer uzet je uvodni citat romana *Beleške jedne Ane* Mome Kapora (1972), koji se može posmatrati u okviru urbane proze namenjene mladima i šireg fenomena tzv. „proze u trapericama” (Flaker, 1983; Bogićević, 2021). Roman karakteriše generacijski obeleženi sleng, realije i elementi fonetike i rime pojedinih idiomatskih izraza koji predstavljaju značajan izazov za generativne modele jer zahtevaju pristup jezičkim informacijama dijahronijske, kulturne i pragmatične prirode (Hendy et al., 2023; Guerreiro et al., 2023).

Dosadašnja istraživanja pokazala su da veliki jezički modeli mogu postići visok stepen fluidnosti u prevođenju, ali da njihova pouzdanost zavisi od jezičkog para, tipologije teksta, kao i dostupnosti jezičko-kulturoloških podataka u obuci modela (Hendy et al., 2023; Guerreiro et al., 2023). Poseban problem predstavljaju jezici koji raspolazu ograničenim digitalnim resursima, kao i tekstovi u kojima se značenje ne zasniva samo na denotativnoj vrednosti leksema, već i na istorijskom, socijalnom, stilskom i kulturološkom kontekstu (Marino, 2022).

Cilj rada je da ispita u kojoj meri različite strategije promptovanja utiču na kvalitet i stabilnost prevoda kulturno i dijahronijski obeleženih književnih tekstova. Pouzdanost dobijenih rezultata biće dodatno proverena ponavljanjem istih promptova u novijoj verziji modela i od strane drugog korisnika. Analizom pojedinačnih prevoda pokušaćemo da odgovorimo na pitanje da li i u kojoj meri tip prompta sistemski determiniše kvalitet generisanih prevoda u obradi jezičkog materijala ove vrste. Polazište za izbor promptova predstavljaju istraživanja o promptovanju u prevođenju, posebno model koji razlikuje osnovni prompt, prevodilački brief i persona-promptove (Gao et al., 2023; He, 2024).

2. Analiza generisanih prevoda

U fokusu analize je uvodni citat romana *Beleške jedne Ane* Mome Kapora, koji je bogat slengom, idiomatskim izrazima, urbanim realijama i fonetskim osobenostima pojedinih izraza (Ljuštanović, 2011; Bogićević, 2021). Izabrani citat nije posmatran kao reprezentativan uzorak celog romana u kvantitativnom smislu, već kao segment

koji omogućava posmatranje ponašanja jezičkog modela u situacijama u kojima prevodilački zadatak prevazilazi puko prenošenje denotativnog značenja reči.

Eksperiment je sproveden u dve faze. U prvoj fazi korišćen je model ChatGPT 5.4 *Thinking*, kome je zadato prevođenje istog odlomka na osnovu četiri različita prompta (tabela 1). Promptovi su oblikovani tako da se razlikuju po stepenu kontekstualizacije i ulozi koja se dodeljuje modelu: od neutralnog zahteva za profesionalnim prevodom, preko prompta sa širim prevodilačkim kontekstom, do persona-promptova u kojima se modelu pripisuje uloga autora popularne proze za mlade, odnosno profesionalnog književnog prevodioca. U drugoj fazi, isti postupak ponovljen je pomoću modela ChatGPT 5.5 *Thinking*, u tri zasebne sesije, a provere je sprovedla i druga korisnica na svom uređaju. Bez obzira na to što time nije formiran statistički reprezentativan korpus generisanih sadržaja, omogućena je provera stabilnosti rezultata, odnosno poređenje početnih prevodilačkih rešenja i kasnijih varijacija dobijenih ponavljanjem istih tipova zadataka.

Tabela 1: Tipovi promptova korišćeni u eksperimentu.

Oznaka prompta	Osnovna karakteristika
<i>Basic</i>	Zahtev za profesionalnim prevodom bez dodatne kontekstualizacije
<i>TransBrief</i>	Prompt sadrži podatke o delu, autoru, godini objavljivanja, ciljnoj publici, urbanoj prozi i prisustvu slenga
<i>Author</i>	Modelu se dodeljuje uloga autora popularne proze za mlade
<i>Translator</i>	Modelu se dodeljuje uloga profesionalnog književnog prevodioca specijalizovanog za prozu za mlade

U analizi nisu uzeti u obzir svi elementi generisanih prevoda, već samo segmenti relevantni za istraživačko pitanje tj. oni koji su leksički, sociolingvistički, pragmatično i kulturno obeleženi. Izdvojene jedinice mogu se podeliti u pet grupa: afektivni i kolokvijalni oblici obraćanja pisca protagonistkinji romana, idiomatske konstrukcije, urbani sleng u dijahroniji, realije gradskog prostora i fonetski elementi i rima u pojedinim izrazima (tabela 2).

Tabela 2: Segmenti izdvojeni za analizu.

<i>Problemska grupa</i>	<i>Segmenti iz izvornog teksta</i>	<i>Prevodilačka rešenja u italijanskom</i>
Afektivni i kolokvijalni oblici obraćanja	vrta repu jedan, zonzaro, folirantkinjo jedna, budalo jedna blesava	„piccola sgualdrina irrequieta”, „piccola civetta”, „zanzara”, „moscona”, „smorfiosa”, „sciocchina svampita”
Idiomatske konstrukcije	bazaš, onako bez veze, za svoj groš; dok ideš tako ničija, bez veze	„vagabondi così, senza una meta, con i tuoi pochi soldi”; „bighelloni così, senza un motivo, per conto tuo”; „di nessuno, senza un perché”
Dijahronijski urbani sleng	mangaši, laponci, genter, peškiri	„teppisti”, „gagà”, „vecchi rimbambiti”, „damerini”, „palpeggiatori”, „cheche”, „tipi loschi”
Realije gradskog prostora	bife, dragstor, robne kuće, gajbe	„buffet”, „bar”, „drugstore”, „grandi magazzini”, „case buie”, „appartamenti bui”, „tane buie”
Fonetski elementi i rima	svi žarko čeznu da te zeznu	„ardono dal desiderio di fregarti”, „bruciano dalla voglia di fregarti”, „muoiono dalla voglia di fregarti”

U prvoj fazi analize pokazalo se da ChatGPT 5.4 *Thinking* uglavnom uspeva da proizvede gramatički i stilski tečan italijanski prevod, ali da se problemi javljaju upravo u segmentima koji su kulturno, dijahronijski i pragmatički obeleženi. Kod afektivnih oblika obraćanja model povremeno uspeva da očuva odgovarajuću konotaciju, kao u rešenjima „piccola civetta” ili „sciocchina svampita”, ali istovremeno nudi rešenja koja menjaju afektivni ton originala. To se

posebno uočava u segmentu „piccola squaldrina irrequieta”, u kome se ton izvornog teksta menja uvođenjem moralnog suda i seksualizacije protagonistkinje, što nije samo stilski neadekvatno, već menja originalni semantički izraz teksta.

Kod idiomatskih izraza model relativno dobro prepoznaje da se u primeru „bazaš onako bez veze“ radi o kretanju bez jasnog cilja, pa se javljaju rešenja poput „vagabondi così”, „bighelloni così” ili „te ne vai in giro così”. Međutim, izraz *za svoj groš* u prvoj fazi često se interpretira doslovno, kao da se odnosi na novac, pa se prevodi kao „con i tuoi pochi soldi”, „coi tuoi quattro soldi” ili „con i tuoi pochi spiccioli” čime se gubi idiomatska vrednost izraza. U kasnijim proverama model povremeno dolazi do jednog od mogućih rešenja „per conto tuo” ili „per i fatti tuoi”, ali se ona ne pojavljuju stabilno u svim proverama.

Najveće poteškoće u prvoj fazi odnose se na dijahronijski urbani sleng. Pojedine lekseme koji se odnose na osobe koje protagonistkinja susreće u svojim lutanjima gradom, model prevodi na osnovu šireg konteksta, pretpostavljajući da oni označavaju opasne, nasilne ili moralno problematične društvene tipove. Tako se *mangaši* koje je u kontekstu moguće povezati sa predstavom urbanih mangupa ili šmekera prevode kao „teppisti”, „bulli” ili „gagà”, pri čemu prva dva rešenja pojačavaju negativnu konotaciju, dok treće uvodi arhaičan italijanski izraz koji nije sasvim odgovarajući na semantičkom nivou. Sličan problem javlja se kod izraza *laponci*, koji se u ovoj prvoj fazi prevodi kao „adescatori”, čime se značenje pomera ka figuri opskurnih tipova loših namera, iako leksikografski podaci¹ upućuju na to da se radi o senilnoj osobi.

Posebno su problematični slučajevi u kojima model nudi prevod koji deluje uverljivo u ciljnom jeziku, dok se u stvarnosti udaljava od izvornog istorijskog i kulturnog konteksta. Takav je slučaj sa izrazom *genteri* za koji ograničeni neformalni izvori² upućuju na moguće nekadašnje značenje policajaca u civilu, a koji se u prvoj fazi prevodi kao „damerini”, što ukazuje na halucinaciju modela koji

¹ Svi primeri su pretraženi u Rečniku srpskoga jezika Matice srpske (2011)

² Radi se o forumu na sajtu Serbiancafe.com u kojem jedan od učesnika izjavljuje: „Ima to u Holivudskim filmovima...to kad ti policajac u civilu pride i pokaže značku...haha! [...] mi smo ih nekad zvali 'genteri'.... a ovi na Zapadu ih zovu undercover police”; <https://www.serbiancafe.com/lat/diskusije/mesg/15/17372517/pascepascenego-sta-ce.html?5>, poslednji pristup 2.4.2026.

rekonstruiše značenje iz konteksta, proizvodeći naizgled moguća, ali ne nužno ispravna rešenja.

Kod izraza *peškiri* uočava se drugačija vrsta oscilacije. U prvoj fazi model nudi neutralizovana ili opšte negativna rešenja, kao što su „tipi loschi” i „tipi viscidi”, čime se gubi specifičnost žargonske aluzije zasnovane na fonetskoj bliskosti leksema „peškir” i pežorativnog naziva za homoseksualce.

Što se tiče realija gradskog prostora izraz *robne kuće* uglavnom se stabilno prevodi kao „grandi magazzini”, dok su *bife* i *dragstor* problematičniji. Prevod „buffet” u italijanskom ne aktivira isti urbani prostor kao srpski bife, koji se u leksikografskim izvorima definiše kao mala ugostiteljska radnja u kojoj se služe pića i hladna jela, kao što ni „drugstore” ne prenosi specifično značenje prodavnice otvorene dvadeset i četiri sata. Kod izraza *gajbe*, koji pored osnovnog značenja ima i žargonsko značenje stana, u prvoj fazi se javljaju prevodi poput „case buie” i „catapecchie buie”. Ovaj primer ne ukazuje samo na ograničenje modela, već i na objektivnu prevodilačku teškoću s obzirom na to da italijanski ne raspolaže žargonskim ekvivalentom koji bi preneo isto značenje. Rešenja poput „case” ili „appartamenti” prenose denotativno značenje, ali neutralizuju registar, dok „catapecchie” i „tane” uvode dodatne konotacije zapuštenosti, opasnosti ili degradacije koje nisu nužno sadržane u izvornom leksemu.

Fonetski nivo i rima ostaju nerešeni u obe faze promptovanja. Izraz *svi žarko čeznu da te zeznu* u svim proverama uglavnom se prevodi semantički korektno, na primer kao „ardono dal desiderio di fregarti”, „bruciano dalla voglia di fregarti” ili „muoiono dalla voglia di fregarti”. Izostaje dakle rima *čeznu/zeznu*, kao i fonetski efekat nastao ponavljanjem frikativa i afrikata. Time se potvrđuje da model u nekim slučajevima relativno lako prenosi semantičku vrednost izraza, ali ne i njegovu fonetsku funkciju i rimu u književnom tekstu, što je u skladu sa nalazima koji ukazuju na razliku između semantičke prihvatljivosti generisanog prevoda i očuvanja stilskih, emotivnih i poetskih slojeva izvornog teksta (Manapbayeva et al., 2024). U drugoj fazi eksperimenta provereno je da li novija verzija modela i ponavljanje zadatka dovode do stabilnijih rešenja. Rezultati pokazuju da ChatGPT 5.5 *Thinking* donosi veću varijabilnost, a u pojedinim slučajevima i preciznije prevodilačke predloge, ali da ta poboljšanja nisu stabilna. Radi preglednosti, rezultati druge faze prikazani su u tabeli 3 u kojoj se navodi vrsta promene u odnosu na početne prevode.

Tabela 3: Stabilnost i promene u dodatnim proverama sa modelom ChatGPT 5.5 *Thinking*.

<i>Problemska grupa</i>	<i>Primeri iz dodatnih provera</i>	<i>Procena stabilnosti</i>
Afektivni i kolokvijalni oblici obraćanja	„moscona”, „codina che scodinzoli”, „piccola sculettatrice”, „piccola imposta”, „pazzerella”, „povera svitata”	Delimično i sporadično poboljšanje (<i>moscona</i>), uz izraženu nestabilnost registra.
Idiomatske konstrukcije	„per conto tuo”, „per i fatti tuoi”, „senza un perché”, „senza motivo”, „tanto per fare” ali i „a tue spese”, „di tasca tua”, „coi tuoi due soldi”	Poboljšanje je moguće, ali nije stabilno.
Dijahronijski urbani sleng	mangaši („spacconi”, „smargiassi”, „ganzi da bar”); laponci („vecchi rimbambiti”, „vecchi bavosi”); gentera („gagà”, „palpeggiatori”, „guardoni”); peškiri („checche”, „froci”, „finocchi”, „invertiti”)	Veća leksička raznovrsnost, pojedina semantička pomeranja („vecchi bavosi”) i nerešene halucinacije („gagà”, „palpeggiatori”, „guardoni”), pragmatički prenaplašeno pežorativni prevodi („froci”, „invertiti”)
Realije gradskog prostora	„buffet”, „bar”, „bettole”, „drugstore”, „grandi magazzini”, „appartamenti bui”, „tane buie”	Delimična stabilnost izraza sa direktnim ekvivalentom („grandi magazzini”), dok se kod lokalno i istorijski obeleženih realija značenje neutralizuje ili pomera (bife: buffet”, „bar”, „bettole”; dragstor: „drugstore”; gajbe: „appartamenti bui”, „tane buie”).
Fonetski elementi i rima	“bramano ardentemente di fregarti”, smaniano dalla voglia di fregarti”, “muoiono dalla voglia di fregarti”, “ardono dal desiderio di fregarti”	Stabilnost semantičke korektnosti; kritični elementi (fonetika, rima) ostaju nerešeni

Rezultati pokazuju da se rezultat promptovanja ne može jednostavno opisati opozicijom uspešno/neuspešno s obzirom na to da dodatni kontekst i dodeljivanje uloge modelu utiču na raznovrsnost leksike, stepen kolokvijalnosti i povremeno približavanje prevoda izvor-nom značenju. U slučaju leksike u dijahroniji, kao i realija, model ni u jednoj fazi ne nudi stabilna rešenja i zadržava eventualne početne halucinacije. Uočava se da različiti promptovi proširuju spektar mogućih rešenja, ali u isto vreme ne nude kriterijum po kome bi se među tim rešenjima mogla pouzdano razlikovati najadekvatnija na stilskom, kulturološkom i dijahronijskom planu.

2.1. Zaključak

Analiza uvodnog citata romana *Beleške jedne Ane* pokazuje da ChatGPT u prevodu sa srpskog na italijanski uglavnom uspeva da proizvede gramatički korektan i čitljiv ciljni tekst, ali da se njegova ograničenja uočavaju u onim segmentima koji su karakteristični za Kaporovu prozu kao što su sleng, idiomatski izrazi, realije i fonetske varijacije izraza. Ponavljanje promptova uticalo je na veći izbor i povremenu veću preciznost pojedinih prevoda, posebno u slučajevima kada model prepoznaje idiomatske i/ili kolokvijalne vrednosti izraza. Pokazalo se međutim da ta rešenja nisu stabilna na semantičkom i konotativnom planu.

U slučaju analiziranog prevoda promptovanje se ne može posmatrati kao postupak koji sam po sebi rešava probleme književnog prevođenja, već pre kao način otvaranja šireg spektra mogućih prevodilačkih rešenja. Konačna procena ostaje u domenu ljudske interpretacije, zasnovane na poznavanju različitih književnih konteksta i stila određenog autora. Književni tekstovi imaju važnu ulogu u proceni kvaliteta prevoda velikih jezičkih modela jer ukazuju na ponašanje modela pred jezičkim materijalom koji je istorijski, kulturno i stilski obeležen. U tom smislu smatramo da bi poboljšanje generativnih modela trebalo biti zasnovano na uključivanju leksikografskih i korpusnih resursa u njihovu obuku i evaluaciju kroz proces interdisciplinarnе saradnje stručnjaka za računarsku lingvistiku i humanističke nauke, posebno u oblasti prevođenja.

References

- Dušan Bogićević. 2021. Duh modernosti u prozi Mome Kapora. *Metodička teorija i praksa*, (2):162–173.
- Aleksandar Flaker. 1983. *Proza u trapericama*. Sveučilišna naklada Liber, Zagreb.
- Yuan Gao, Rui Wang, and Feng Hou. 2023. [How to Design Translation Prompts for ChatGPT: An Empirical Study](#). *arXiv preprint arXiv:2304.02182*.
- Nuno M. Guerreiro, Duarte M. Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André F. T. Martins. 2023. [Hallucinations in Large Multilingual Translation Models](#). *Transactions of the Association for Computational Linguistics*, 11:1500–1517.
- Shuang He. 2024. [Prompting ChatGPT for Translation: A Comparative Analysis of Translation Brief and Persona Prompts](#). In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation*, volume 1, pages 316–326, Sheffield, UK. European Association for Machine Translation.
- Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. [How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation](#). *arXiv preprint arXiv:2302.09210*.
- Momo Kapor. 1972. *Beleške jedne Ane*. Beograd.
- Jovan Ljuštanović. 2011. Momo Kapor: Beleške jedne Ane – prostor i vreme. In *Pripovedač urbane melanholiije: književno delo Mome Kapora*, pages 111–125. Učiteljski fakultet, Beograd.
- Zhannura Manapbayeva, Gulkhadisha Zaurbekova, Karlygash Ayazbekova, Aigul Kazezova, and Kunsulu Pirmanova. 2024. [AI in Literary Translation: ChatGPT-4 vs. Professional Human Translation of Abai’s Poem “Spring”](#). *Procedia Computer Science*, 251:526–531.
- Stefano Marino. 2022. [La \(ragionevole\) prassi dialettica della traduzione e le sue regole](#). *Studi di estetica*, 50(1):95–116.
- Milica Vujanić, Darinka Gortan-Premk, Milorad Dešić, Rajna Dragićević, Miroslav Nikolić, Ljiljana Nogo, Vasa Pavković, Nikola Ramić,

Rada Stijović, Milica Radović-Tešić, Egon Fekete, and Miroslav Nikolić. 2011. *Rečnik srpskoga jezika*. Matica srpska, Novi Sad.

Momo Kapor in the Digital Environment: Literary Translation and ChatGPT

Olja Perišić

Summary

This paper examines the capabilities and limitations of ChatGPT in translating literary texts from Serbian into Italian, using the opening passage of Momo Kapor's novel *Beleške jedne Ane* (Ana's Notes, 1972) as a case study. The novel belongs to the tradition of urban youth prose and the broader literary phenomenon known as "jeans prose". Its language is marked by generational slang, idiomatic expressions, culture-specific references to Belgrade in the 1970s, colloquial forms of address, and phonetic and rhyming effects. These features pose a particular challenge to generative translation based on large language models because their interpretation requires not only linguistic competence but also diachronic, sociocultural, pragmatic, and stylistic knowledge.

The aim of the study is to determine how different prompting strategies affect the quality and stability of translations of culturally, historically, and stylistically marked literary language. The experiment was conducted in two stages. In the first stage, the same passage was translated using ChatGPT 5.4 *Thinking* and four prompt types: a basic prompt requesting a professional translation without additional contextualization; a translation-brief prompt containing information about the novel, its author, year of publication, intended readership, urban setting, and use of slang; and two persona prompts assigning the model the role of either an author of popular fiction for young readers or a professional literary translator specializing in youth literature. In the second stage, the procedure was repeated with ChatGPT 5.5 *Thinking* in several separate sessions and by different users in order to assess the stability of the generated solutions.

The analysis focuses on five groups of translation problems: affective and colloquial forms of address, idiomatic constructions, diachronic urban slang, culture-specific elements of the urban environment, and phonetic features and rhyme. The results show that ChatGPT generally produces grammatically correct, fluent, and apparently coherent Italian texts. Nevertheless, its most significant limitations emerge precisely in passages whose interpretation depends on historical, cultural, or pragmatic context. The model occasionally changes the affective tone of the original, interprets idiomatic expressions literally, neutralizes register-specific meanings, or introduces connotations that are absent from the source text. Diachronic

slang proves particularly problematic: the proposed Italian equivalents are often semantically unstable, excessively pejorative, or based on plausible but unsupported interpretations. Such solutions may sound convincing in the target language while departing substantially from the meaning and cultural context of the original.

Repeated prompting and the use of a newer model increase lexical diversity and occasionally lead to more accurate translations. However, these improvements are neither systematic nor stable across sessions and users. Additional contextual information and persona prompts broaden the range of possible solutions but do not provide a reliable criterion for selecting the most appropriate equivalent. Moreover, although the semantic content of expressions involving sound patterning is generally conveyed, their rhyme and phonetic effects remain unresolved.

The findings suggest that prompting alone cannot overcome the central difficulties of literary translation. It can support translators by generating alternative formulations, but the final assessment still requires human interpretation, knowledge of the author's style, and familiarity with the relevant linguistic and cultural context. Improving large language models for literary translation, particularly for languages with fewer digital resources such as Serbian, therefore requires the systematic integration of diachronic dictionaries, specialized corpora, and culturally informed lexical resources into model adaptation and evaluation, supported by interdisciplinary collaboration between computational linguists, literary scholars, and translation experts.

Keywords: ChatGPT, literary translation, Momo Kapor

Works are licensed under the terms and conditions of the Creative Commons Attribution 4.0 International License.

<https://creativecommons.org/licenses/by/4.0/>



Combining Scholarly Edition and NLP: Historical Multilingual Parallel Text Collections on TEITOK

Maarten Janssen¹[0000-0003-0272-6318], Anna Jouravel²[0000-0002-7767-4973],
Piroska Lendvai³[0000-0001-6670-9475]

¹Charles University, ÚFAL, Faculty of Mathematics and Physics, Czechia,
janssen@ufal.mff.cuni.cz

²University of Freiburg, Department of Slavic Languages and Literatures,
Germany, anna.jouravel@slavistik.uni-freiburg.de

³Bavarian Academy of Sciences and Humanities, Department of Digital
Humanities, Germany, piroska.lendvai@badw.de

Abstract

The proposed presentation demonstrates a workflow for transforming a semi-digital scholarly edition of a medieval text into a multilingual historical NLP resource (Janssen et al., 2026). Our case study is *De lepra ad Sistelium* by Methodius of Olympus, a 4th-century Greek treatise preserved only fragmentarily in Greek but more fully in medieval Slavic translation. Starting from a recent semi-digital edition and additional digitized witnesses, diplomatic transcriptions, and modern translations, we compile a parallel collection on the TEITOK platform. The paper focuses on text-to-text alignment in the philological sense, that is, the modelling of correspondences between passages, sentences, and tokens across heterogeneous textual witnesses. All sources are converted into TEI/XML and kept as richly annotated documents. Alignment is represented through shared translation-unit identifiers (@tuid) on XML nodes, which support one-to-one, one-to-many, and many-to-many correspondences and are therefore suited to historical material with divergent segmentation, omissions, expansions, and other non-isomorphic relations. We also describe morphosyntactic enrichment with UDPipe and subsequent manual correction for Ancient Greek and Church Slavonic. The resulting collection is both a searchable multilingual corpus and an emerging gold standard for low-resource historical NLP.

Keywords: historical NLP, TEITOK, text alignment, Church Slavonic, TEI/XML, digital philology

1. Introduction

The increasing availability of digital scholarly editions, manuscript transcriptions, and born-digital translations creates new possibilities for the computational study of pre-modern textual traditions. At the same time, such materials pose difficulties that are less prominent in modern NLP pipelines. Historical witnesses are rarely plain text. They are usually embedded in a rich editorial structure and frequently display orthographic variation, unstable punctuation, omissions, additions, or non-corresponding segmentation. Historical source materials often contain rich editorial and document-structural markup, including variant readings, layout information, and multiple layers of linguistic interpretation. This kind of TEI/XML representation is highly valuable for philological work, but it can be difficult to integrate directly into standard NLP pipelines. As a result, some workflows first convert TEI-encoded sources into simpler processing formats such as plain text, pseudo-CoNLL, or project-specific XML in order to enable automatic annotation or alignment (Basile and Sangati, 2016; Vaamonde, 2018; Lavrentiev, 2012). Our approach follows a different logic: we keep alignment and annotation anchored in TEI/XML throughout. In this way, editorial structure and linguistic enrichment remain connected within the same representation, which is especially important for historical material with non-isomorphic segmentation, variant transmission, and complex document structure.

This contribution presents such a workflow on the basis of the treatise *De lepra ad Sistelium* by Methodius of Olympus. The text was composed in Greek at the turn of the 4th century and offers an allegorical interpretation of the Old Testament regulations on leprosy (Lev 13). Its original Greek text is preserved only fragmentarily, whereas a much fuller 10th-century Church Slavonic translation has survived in later manuscript copies (Jouravel and Sieber, 2024; Maksimczuk, 2024). This asymmetrical transmission makes the text a particularly instructive case for text-to-text alignment: while the correspondences between witnesses are often philologically clear, they do not always align at the sentence level because segmentation,

omissions, expansions, and reordering may differ across versions and sentence boundaries in a modern sense are not indicated.

Our aim is therefore twofold. First, we show how heterogeneous source materials can be transformed into a common TEI/XML representation and integrated into the TEITOK platform (Janssen, 2016). Secondly, we argue that the resulting resource goes beyond the representation mode a digital edition offers. By combining TEI-based alignment, morphosyntactic enrichment, and corpus search, the workflow yields a reusable historical NLP resource that is of interest to philology, corpus linguistics, and multilingual NLP alike.¹ Building on our earlier discussion of automatic alignment methods for this material (Janssen et al., 2025, 2026), we now shift the emphasis from tool evaluation to corpus design and scholarly reuse.

2. Source material and corpus design

The collection is based on the recent semi-digital edition of *De lepra* (Jouravel et al., 2024; Bracht et al., 2024) prepared in the *Classical Text Editor* (Hagel, 2007). From this edition, we derived five textual representations: the reconstructed Greek text, a German translation of the Greek, the reconstructed Slavonic text, a German translation of the Slavonic, and a diplomatic transcription of the 16th

¹ For (medieval) Slavic Studies, we are currently aware of only one other project that combines a scholarly edition with a corpus-like aligned presentation environment, namely the online *Dioptra* edition (Fuchsbaauer and Maion, 2025). The interface is designed primarily for the parallel display of Greek and Slavonic, but the HTML structure also contains visible placeholders for a German layer, even though this does not appear to be activated (yet?) in the public reading interface. It offers a reading and a search mode, lets the user switch displayed versions on and off, highlights commentary, and exposes lemma-based search and some syntactic search options. There is also a statistics panel for tokens, types, and lemmata, and a placeholder for syntax-tree display. Technically, the alignment seems to be implemented directly in the web presentation layer: corresponding Greek and Slavonic tokens are rendered as HTML spans that share the same identifier, while some spans carry grouped identifiers, which suggests that one displayed element may be linked to several positions. Additional lexical and editorial information is attached to the same spans through attributes such as lemma and comment fields. What remains unclear from the public interface alone is how this alignment was originally produced. The page source shows the result of an alignment, but not its workflow history; we therefore cannot determine with confidence whether the linking was established manually, semi-manually, or by automatic pre-processing followed by correction.

century manuscript used for the Slavonic edition.² To these we added a second diplomatic Slavonic transcription as preserved in the 16th century *Great Menaion Reader* as well as an earlier German (Bonwetsch, 1917) and English (Pearse et al., 2015) translation of the treatise. The result is a small but structurally rich corpus that combines editions, diplomatic transcriptions, and modern translations.

The linguistic status of the Slavonic witnesses deserves a brief note. Although the translation ultimately goes back to the (Old) Church Slavonic milieu (OCS), the surviving copies belong to an Old East Slavonic (OES) recension. For corpus annotation we therefore treat the material pragmatically as *orv*, the language code for OES, since existing taggers trained on OES data perform more reliably on the orthography of these witnesses than taggers targeting canonical OCS (Berdicevskis and Eckhoff, 2020).

Table 1 gives an overview of the text’s different textual and editorial representations used in the aligned collection.

Table 1: Source material used in the TEITOK collection.

ID	Lang.	Description	Sentences	Tokens
A	grc	Greek critical edition	52	1762
B	deu	German translation of A	213	2225
C	orv	Slavonic edition	267	5068
D	deu	German translation of C	270	6552
E	orv	Diplomatic transcription of manuscript used for C	702	4561
F	orv	Diplomatic transcription of Great Menaion Reader witness	493	4596
G	deu	Earlier German translation	241	6584
H	eng	English translation	298	9076

All these representations may be parallelized; this makes it possible to compare, for example, a normalized Church Slavonic edition with diplomatic transcriptions, or to relate different modern translations to the same historical passage. The resource can thus be used to investigate how corresponding content is segmented, expanded, omitted, or reformulated across witnesses and translations. On the other hand, the morphosyntactic annotation supports linguistic analysis.

² It is the manuscript RNB Q.I. 265, a volume kept in the Russian National Library in St. Petersburg containing only the author’s works.

3. Conversion and alignment modelling

All source materials were converted into TEI/XML before ingestion into TEITOK.³ Part of the corpus was already available in TEI-like form through the *Classical Text Editor* export; other parts had to be converted from HTML or PageXML and subsequently normalized for TEITOK. We preserve editorial information relevant for later visualization and analysis within the TEI structure. Figure 1 shows the TEITOK alignment view used for inspecting and correcting correspondences between textual witnesses.

Alignment viewer

File: awesome-82d.json
 Version 1: cte-cho-trans.xml
 Version 2: delepra-cho-deu.xml
 Method: Awesome-align on raw XML
 Script: awesome.py

ID version 1	Text version 1	Gold comparison	ID version 2	Text version 2
p-42	Where are you from, O Eubulus?	p-6	p-6	Woher kommet du, Eubulus?
p-47	Sit.	p-4	p-4	Sit.
p-43	Is it not clear that from the path, taking according to individual understanding of divine books?	p-7	p-7	Ist es denn nicht offensichtlich, dass von der Jagdnach der Verständnisweise der göttlichen Schrift?
p-43a	Eub.	p-6	p-6	Eub.
p-48	And Sistelus said: 'Now, let us look at the truth itself,' the books; we will see that evil is driven out by the words of healing, as if by a medicine, saying: 'Arise, and Awake from the dead, and Christ will enlighten you.'	p-10	p-11a 12a-9	Und Sistelus sprach: „Nun können wir nämlich aus der Wahrheit der Schrift selbst ersehnen diesen schlimmen Aussatz mit Heilkraut vertreiben, indem wir mit Worten, wie mit Medizin den Schaden lindernd, sagen: «Erhebe dich, der du schläfst, und stehe auf von den Toten, und Christus wird dich erleuchten.» Jetzt ist es aber an der Zeit, dass du das Wort ergreifst“, sagte er „Wohlan denn, erläutere uns den Aussatz, der im Gesetz steht, denn der Herr legt nahe, vieles allegorisch auszulegen, indem er im Gesetz dazu auffordert, demjenigen eine reichliche Gabe mitzuteilen, der nach Gottes Gabe dürstet.“ Und als wir uns seinem Hof näherten, kam Sistelus mir entgegen, umarmte mich und sprach: „Seit du von uns gegangen bist, hast du uns des Schriftverständnisses beraubt; denn wie eine Wolke, die sich vor die Sonne stellt, nicht gestattet, die Sonne klar zu sehen, so sind auch, wo sich die rechte Unterweisung entfernt, die Seelen dunkel und der Verstand verfinstert.“ Jetzt ist es aber an der Zeit, dass du das Wort ergreifst“, sagte er „Wohlan denn, erläutere uns den Aussatz, der im Gesetz steht, denn der Herr legt nahe, vieles allegorisch auszulegen, indem er im Gesetz dazu auffordert, demjenigen eine reichliche Gabe mitzuteilen, der nach Gottes Gabe dürstet.“ Und Sistelus sprach: „Nun können wir nämlich aus der Wahrheit der Schrift selbst ersehnen diesen schlimmen Aussatz mit Heilkraut vertreiben, indem wir mit Worten, wie mit Medizin den Schaden lindernd, sagen: «Erhebe dich, der du schläfst, und stehe auf von den Toten, und Christus wird dich erleuchten.» Und ich antwortete ihm: „Das hast du schön gesagt!“ Und wir gingen ins Haus und setzten uns.
p-50	Go now and speak about what is written in the law concerning leprosy, which the Lord commands to be similar to the gift of God's generosity to those who desire it.	p-12	p-12a 11a-10	
p-44	(For in the morning, someone knocked on the door of Sistelus, and a servant opened it for him and led him to call Sistelus to himself, and rising, I went there immediately.)	p-8	p-8	(In der Frühe klopfte nämlich jemand von den Leuten des Sistelus an die Tür, und als ein Diener ihm öffnete und berichtete, Sistelus rufe uns zu sich, stand ich auf und ging sogleich hin.

Figure 1: Alignment view in TEITOK.

The key methodological choice is the representation of alignment by means of shared translation-unit identifiers (@tuid) assigned directly to XML nodes. Two nodes in different documents are aligned if they share the same identifier. Because these identifiers can be attached to different levels of the document hierarchy, alignment can be represented simultaneously at the level of text, section, paragraph, sentence, and token. More importantly, a node may contain a list of identifiers. This makes it possible to represent one-to-many, many-to-one, and many-to-many correspondences. Such flexibility is indispensable for historical witnesses, preserved to varying degrees of

³ The collection can be accessed online: https://lindat.mff.cuni.cz/services/teitok-live/methodius_lepra/index.php?action=files&folder=xmldatafiles/delepra.

completeness and their modern translations, that sometimes regularize punctuation.

This many-to-many model distinguishes our approach from stricter alignment formats such as TMX, which assume a single alignment level and typically require a one-to-one mapping between translation units. It also differs from pairwise alignment formats, which can be expressive but become cumbersome when several textual witnesses need to be maintained simultaneously. In TEITOK, the `@tuid`-based solution allows us to keep each witness as a full TEI/XML document while generating synoptic display dynamically from the alignment links. Alignment is therefore not stored in a separate table detached from the source representation, but embedded in the textual structure itself (Janssen, 2016; Janssen et al., 2025).

4. Automatic pre-alignment and manual correction

The aligned collection was produced semi-automatically. For each new witness, we first generated an automatic bitext alignment in JSON and then manually corrected it in the TEITOK environment before converting it into the final `@tuid`-based representation.

Our evaluation of automatic alignment methods on the *De lepra* data has shown that historical materials remain difficult for current alignment tools (Janssen et al., 2025). We tested Hunalign (Varga et al., 2007), LaBSE (Feng et al., 2022), and Awesome-Align (Dou and Neubig, 2021). The results were instructive: despite the current prominence of embedding-based and LLM-related approaches, Hunalign performed better overall on this dataset than the neural methods.⁴ The corpus includes omissions, abbreviated passages, major differences in segmentation, and witnesses that should often remain partially unaligned. For this TEITOK provides a dedicated alignment editor in which misaligned units can be reassigned manually. Once a witness has been corrected, its alignments become available for the whole collection through the shared `@tuid` system.

5. Morphosyntactic enrichment and historical NLP

Beyond textual alignment, the corpus was enriched with morphosyntactic annotation using *UDPipe* (Straka, 2018). We ap-

⁴ The reason is not that historical data are inherently resistant to modern models, but that the relation between our witnesses differs substantially from the type of parallel data on which such models are usually trained.

plied lemmatization, part-of-speech tagging, and dependency parsing, using modern English and German models for the translations and historical models based on PROIEL and TOROT for Greek and Old East Slavonic (Haug and Jøhndal, 2008; Eckhoff et al., 2017). For the historical languages, however, the output required extensive manual correction,⁵ which – although being labour-intensive – created new gold-standard data for medieval Slavonic and Greek material, which remain underrepresented in current NLP resources.

An additional advantage of the parallel design is annotation transfer. Where closely related witnesses exist, corrected annotation can be projected from one witness to another. This does not eliminate the need for expert supervision and manual correction where needed, but it reduces redundant work and shows how alignment and annotation can reinforce one another in a historically oriented corpus workflow.

6. Visualization, search, and scholarly use

A major strength of the TEITOK implementation lies in the fact that – beside storage – the aligned data are made explorable in several complementary ways. First, the platform offers a parallel running-text view in which full TEI documents can be displayed side by side. Hovering over a segment highlights the corresponding passages in the other witnesses.

Secondly, TEITOK can generate a table of aligned elements. This view is suited to local comparison, for instance when inspecting sentence-level correspondences across several versions.

Thirdly, the resource is fully searchable. Standard corpus queries can target forms, lemmas, or parts of speech using the integrated corpus query engine (Evert and Hardie, 2011). Crucially, search results remain linked to the alignment structure. From any hit, the user can open the full context in the relevant document or display all witnesses connected to the same translation unit. This makes the corpus useful for several communities: philologists can investigate transmission variants and translation choices; corpus linguists can explore multilingual annotation; and NLP researchers can work with a richly structured aligned dataset.

⁵ The gap between training data and the mediaeval orthographic as well as modern transcriptional conventions of our sources remains substantial.

Future work will expand the corpus with additional witnesses and further layers of the scholarly edition, including apparatus material. Just as importantly, the corrected annotations and alignment links can serve as training and evaluation data for future models of historical text processing. In this sense, the work neither treats philology as a preliminary step before NLP, nor sees NLP as an add-on to editorial work, but develops an infrastructure in which both are mutually informative.

Acknowledgements

This work was partly funded by the German Federal Ministry of Research, Technology and Space (BMFTR; "next generation EU"), and was part of the collaboration between the *QuantiSlav* project (University of Freiburg and Bavarian Academy of Sciences and Humanities, Munich) and the TEITOK platform developer (Charles University, Prague).

References

- Angelo Basile and Federico Sangati. 2016. *D(H)ante: A New Set of Tools for XIII Century Italian*. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2825–2828, Portorož, Slovenia. European Language Resources Association (ELRA).
- Aleksandrs Berdicevskis and Hanne Eckhoff. 2020. *A Diachronic Treebank of Russian Spanning More Than a Thousand Years*. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5251–5256, Marseille, France. European Language Resources Association.
- Gottlieb Nathanael Bonwetsch. 1917. *Methodius*. J.C. Hinrich, Leipzig.
- Katharina Bracht, Anna Jouravel, and Janina Sieber, editors. 2024. *Methodius of Olympos: De lepra*. De Gruyter, Berlin, Boston.
- Zi-Yi Dou and Graham Neubig. 2021. *Word Alignment by Fine-tuning Embeddings on Parallel Corpora*. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128. Association for Computational Linguistics.

- Hanne Eckhoff, Kristin Bech, Gosse Bouma, Kristin Eide, Dag Trygve Truslew Haug, Øyvind B. Haugen, and Marius L. Jøhndal. 2017. [The PROIEL treebank family: a standard for early attestations of Indo-European languages](#). *Language Resources and Evaluation*, 52(1):29–65.
- Stefan Evert and Andrew Hardie. 2011. [Twenty-First Century Corpus Workbench: Updating a Query Architecture for the New Millennium](#). In *Corpus Linguistics 2011*.
- Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. [Language-agnostic BERT Sentence Embedding](#).
- Jürgen Fuchsbauer and Fabio Maion. 2025. [Dioptra Online: Eine morphologisch annotierte digitale Edition des byzantinischen Textes und seiner slawischen Übersetzung](#).
- Stefan Hagel. 2007. [The Classical Text Editor. An Attempt to Provide for Both Printed and Digital Editions](#). In A. Ciula and F. Stella, editors, *Digital Philology and Medieval Texts*, pages 77–84. Pacini Editore, Pisa.
- Dag Trygve Truslew Haug and Marius L. Jøhndal. 2008. [Creating a Parallel Treebank of the Old Indo-European BibleTranslations](#).
- Maarten Janssen. 2016. [TEITOK: Text-faithful annotated corpora](#). *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016*, pages 4037–4043.
- Maarten Janssen, Anna Jouravel, and Piroska Lendvai. 2026. [From Semi-Digital Edition to Historical NLP Resource:Constructing and Annotating Historical Multilingual Parallel Text Collections on the TEITOK Platform](#). In *The Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 1553–1561, Palma, Mallorca, Spain.
- Maarten Janssen, Piroska Lendvai, and Anna Jouravel. 2025. [Alignment of Historical Manuscript Transcriptions and Translations](#). In *Proc. of RANLP 2025, September 2025, Varna, Bulgaria*.
- Anna Jouravel and Janina Sieber. 2024. [The Greek and Slavonic Transmission of Methodius’ De lepra](#). In Katharina Bracht, Anna Jouravel, and Janina Sieber, editors, *Methodius of Olympus: De lepra*, pages 11–30. De Gruyter, Berlin, Boston.
- Anna Jouravel, Janina Sieber, and Katharina Bracht. 2024. *Methodius von Olympus: De lepra. Griechischer und slavischer Text*,

mit Einleitung und deutscher Übersetzung., volume NF, 31 of *Die griechischen christlichen Schriftsteller der ersten drei Jahrhunderte*. De Gruyter, Berlin.

- Aleksei Mikhailovich Lavrentiev. 2012. Проблемы лингвистической разметки и анализа электронных критических изданий текстов письменного наследия в стандарте XML-TEI.
- José Maksimczuk. 2024. The Florilegium Coislinianum and the Greek Text of Methodius’ *De lepra*. In Katharina Bracht, Anna Jouravel, and Janina Sieber, editors, *Methodius of Olympus: De lepra*, pages 31–54. De Gruyter, Berlin, Boston.
- Roger Pearce, Ralph Cleminson, and Andrew Eastbourne. 2015. *Methodius of Olympus, De Lepra (On Leprosy)*. Roger Pearce.
- Milan Straka. 2018. UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task. In *Proceedings of The*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.
- Gael Vaamonde. 2018. La multidisciplinariedad en la creación de corpus históricos: El caso de Post Scriptum. *Artnodes*.
- Dániel Varga, Péter Halácsy, András Kornai, Viktor Nagy, László Németh, and Viktor Trón. 2007. Parallel corpora for medium density languages. In Nicolas Nicolov, Kalina Bontcheva, Galia Angelova, and Ruslan Mitkov, editors, *Selected papers from RANLP 2005*, pages 247–258. John Benjamins Publishing Company.

Повезивање научног издања и NLP-а: историјске вишејезичне паралелне текстуалне колекције на платформи TEITOK

Maarten Janssen, Anna Jouravel, Piroska Lendvai

Апстракт

Овај проширени апстракт представља поступак којим се полудигитално научно издање средњовековног текста претвара у вишејезични историјски NLP ресурс. Као студија случаја узима се спис *De lepra ad Sistelium* Методија Олимпског, грчки трактат из четвртог века који је на грчком сачуван само фрагментарно, док је у средњовековном словенском преводу очуван у знатно потпунијем облику. Полазећи од новијег полудигиталног издања и додатних дигитализованих сведока, дипломатских транскрипција и савремених превода, рад описује формирање паралелне колекције на платформи TEITOK. Тежиште је на текстуалном поравнању у филолошком смислу, односно на моделовању подударности између одломака, реченица и токена у различитим текстуалним сведоцима, транскрипцијама и преводима. Сви извори се конвертују у TEI/XML и остају сачувани као богато аотирани документи. Поравнање се представља помоћу заједничких идентификатора преводних јединица (@tuid) у XML чворовима, што омогућава односе један-према-један, један-према-више и више-према-више, неопходне за историјску грађу са различитом сегментацијом, изостављањима, проширењима и другим несиметричним односима. Колекција је додатно обogaћена морфосинтаксичком аотацијом помоћу UDPipe-а и ручном корекцијом за старогрчки и црквенословенски. Тако настали ресурс функционише истовремено као претражив вишејезични корпус, као окружење за филолошко поређење сведока и као настајући златни стандард за слабо ресурсне историјске језике.

Кључне речи: историјски NLP, TEITOK, текстуално поравнање, црквенословенски, TEI/XML, дигитална филологија

Works are licensed under the terms and conditions of the Creative Commons Attribution 4.0 International License.

<https://creativecommons.org/licenses/by/4.0/>



Језичка анализа Q&A комуникације у јавним позивима за финансирање научноистраживачких пројеката

Ђурђа Перовић¹

¹ Универзитет у Београду, Мултидисциплинарне мастер студије, Рачунарство у друштвеним наукама, Србија, dp.0053.20@student.rect.bg.ac.rs

Апстракт

Студија испитује језичке одлике институционалне комуникације у Q&A системима који прате јавне позиве за финансирање научноистраживачких пројеката. Ови системи представљају структуриран комуникациони канал између апликаната и институције током процеса пријаве научних пројеката.

Истраживање се заснива на корпусу формираном на основу јавно доступне документације Фонда за науку Републике Србије, који обухвата 4.124 пара питања научника и одговарајућих одговора институције, са укупно 494.288 токена, током 9 јавних позива. Применом метода корпусне лингвистике, текстометрије и обраде природног језика (*Natural Language Processing, NLP*), проучавају се разлике између питања и одговора у погледу њихове дужине и структуре, лексичке разноврсности, граматичких и синтаксичких одлика, као и употребе карактеристичних лексема и конструкција. Посебна пажња је посвећена језичким средствима која одликују административни функционални стил и бирократски језик, укључујући номинализацију, карактеристичне предлошке и модалне конструкције.

Добијени резултати показују јасне разлике између језика питања научника и институционалних одговора. Поједини налази додатно се разматрају у поређењу са подацима из референтног корпуса српског језика (SrpKor) и корпуса припремљеног на основу јавних нормативних докумената Фонда за науку Републике Србије који дефинишу услове и процедуре пријаве на дате јавне позиве.

Примена рачунарских метода у анализи овако структурираног корпуса омогућава прецизније сагледавање специфичних језичких образаца који одликују институционални дискурс у односу на језик питања апликаната. Налази могу допринети бољем разумевању званичних упутстава и послужити као основа за унапређење институционалне комуникације.

Кључне речи: институционални дискурс, бирократски језик, корпусна лингвистика, текстометрија, обрада природног језика, научно финансирање

1. Увод

У савременим системима научног финансирања, дигитални канали комуникације између институција и истраживача имају важну улогу у информисању истраживача током конкурсисања за научноистраживачке пројекте. Посебан облик ове комуникације представљају системи за одговарање на питања (Q&A системи), у којима се одговори институције на питања апликаната јавно објављују ради појашњења услова конкурса, правила програма и процедура пријављивања. Међутим, сложеност језика институција може отежати разумевање званичних информација и упутстава (Ahuja and Wadhvani, 2025).

Поред своје основне намене, ови системи истовремено представљају значајан извор језичких података који омогућава контрастивну анализу питања научника и институционалних одговора. Док су одговори институције по правилу формалнији и припадају административном функционалном стилу, питања апликаната су често мање формална и ближа природној језичкој употреби. Овај контраст отвара могућност да се испита како улога учесника у комуникацији обликује језички израз и које су особине карактеристичне за институционалне одговоре.

За потребе овог истраживања формиран је корпус питања и одговора из Q&A система Фонда за науку Републике Србије. Овај корпус повезује питања научника са одговорима институције и тако омогућава њихово непосредно поређење применом рачунарских метода анализе језика.

Циљ рада је да утврди разлике између језика питања и институционалних одговора и да издвоји језичке обрасце карактеристичне за институционалне одговоре. Посебна пажња посвећена је језичким појавама које одликују административни функционални стил и бирократски језик.

2. Теоријски оквир

Институционални дискурс представља специфичан облик јавне комуникације чија је основна функција саопштавање (Тошовић, 2002), односно преношење информација, правила и процеду-

ра у оквиру одређеног друштвеног или административног система. Његове особине произилазе из контекста у којем се комуникација одвија, али и односа између учесника (Drew and Heritage, 1992). У институционалним контекстима учесници имају различите улоге, при чему институција има превасходно нормативну улогу. Такав однос се може одразити и на избор језичких средстава, начин обликовања информација и ниво формалности комуникације. Језик тако истовремено преноси информације и одражава друштвене улоге у комуникацији (Fairclough, 1989; Klikovac, 2008).

У проучавању употребе језика у различитим комуникационим ситуацијама користе се појмови *дискурс*, *регистар* и *функционални стил*, који се у различитим лингвистичким традицијама делимично преклапају, али немају сасвим једнако значење. Док се у лингвистици енглеског језика појам регистар односи на језичке обрасце повезане са одређеном ситуацијом и комуникативном функцијом (Biber and Conrad, 2009), у српској литератури се функционално раслојавање језика описује кроз *функционалне стилове* (Тошовић, 2002). Термин *институционални дискурс* у овом раду користимо у ширем смислу, за језичку комуникацију која се одвија у институционалном контексту, док *административни функционални стил* означава језичке особине карактеристичне за административну комуникацију.

Административни функционални стил обухвата комуникацију која прати рад институција и одликују га обезличеност, тежња ка неекспресивном, прецизном и стандардизованом изражавању, са пуно понављања и глаголских именица (Тошовић, 2002). У оквиру такве комуникације могу се јавити особине које се повезују са бирократизацијом језика. Кликовац (2001: 87) издваја: *номиналност*, *општост*, *неодређеност*, *експлицитност* у смислу редунадантности, *еуфемизам*, *вишак речи* и *квазинаучност*, као и *употребу речи страног порекла* и *помодних речи*. Неке од ових појава, попут номиналности и експлицитности, релевантне су и за анализу институционалних одговора у овом истраживању.

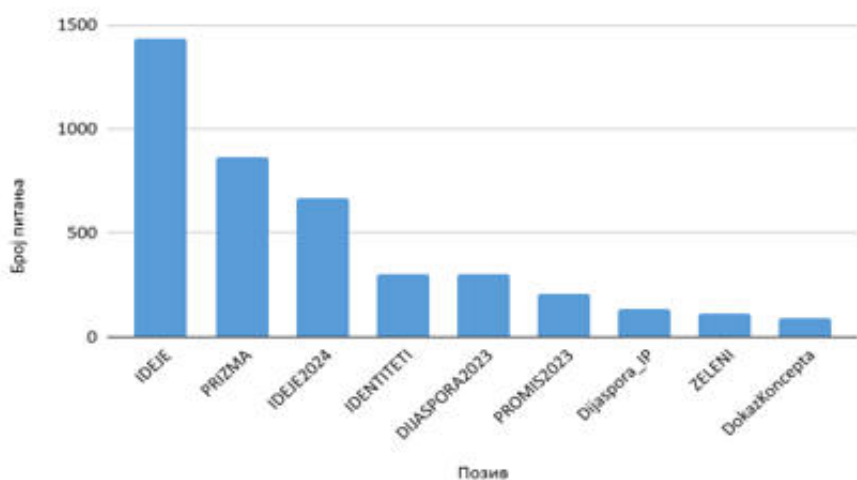
Корпусни приступ омогућава да се језичке појаве испитују на основу већег броја текстова и ефикасније пореде између различитих делова корпуса (McEnery and Hardie, 2012). Захваљујући рачунарским алатима, могуће је брзо обрадити велику количину текста и издвојити обрасце који нису увек лако уочљиви

ви приликом читања. Такав приступ у овом раду примећујемо на питања научника и одговора институције, како бисмо прецизније сагледали језичке разлике и обрасце карактеристичне за ову врсту институционалне комуникације.

3. Развој корпуса

За потребе рада формирана су два корпуса: 1) корпус питања и одговора (Q&A парова) и 2) нормативни корпус.

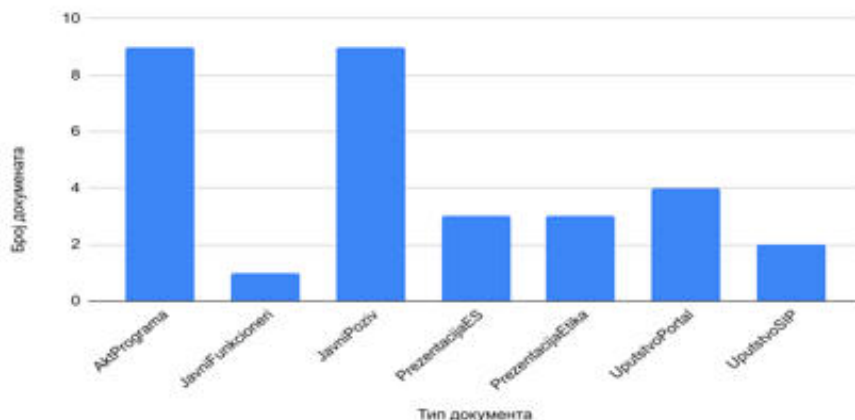
Корпус питања и одговора обухвата питања истраживача и одговоре институције објављене на званичном сајту Фонда за науку Републике Србије у периоду од 2020. до 2025. године, у оквиру 9 јавних позива за финансирање научноистраживачких пројеката (Слика 1). Састоји се од 4.124 Q&A пара, прикупљена из 62 јавно доступна Q&A документа. Подаци су организовани у два поткорпуса, питања апликаната и институционалне одговоре, и обogaћени метаподацима о извору, години, јавном позиву и теми питања, на основу које су парови сврстани у 42 тематске категорије.



Слика 1: Број Q&A парова по јавним позивима у корпусу питања и одговора.

Нормативни корпус је формиран на основу 40 јавно доступних докумената који дефинишу правила и процедуре пријављивања на јавне позиве Фонда за науку Републике Србије (Слика 2)

(Фонд за науку РС, 2020–2025). Обухвата различите типове конкурсне документације и коришћен је као референтни корпус за поређење појединих језичких појава.



Слика 2: Типови докумената у нормативном корпусу.

Захваљујући структури упарених питања и одговора, корпус омогућава непосредно поређење језика апликаната и институције. Подаци су структурирани и припремљени у машински читљивом формату, што омогућава њихову рачунарску обраду и текстометријску анализу.

4. Методологија и анализа корпуса

Сprovedена је квантитативно-квалитативна анализа корпуса питања и одговора, који је подељен на поткорпус питања апликаната и поткорпус институционалних одговора. Анализа је обухватила дужину и структуру текстова, лексичку разноврсност, као и издвајање лексичких јединица и конструкција карактеристичних за поједине делове корпуса. У испитивању њихових језичких разлика, комбиноване су текстометријске методе и методе обраде природног језика.

За претрагу, израду конкорданци и статистичку обраду корпуса коришћен је програм TXM (Heiden, 2010), уз интегрисани алат TreeTagger и модел SrpKor4Tagging за аутоматско одређивање лема и врста речи у српском језику (Stanković et al., 2020). За обраду и анализу података коришћен је програмски језик Python,

помоћу ког су израчунати показатељи дужине питања и одговора и њихових односа, као и мере лексичке разноврсности.

Према подацима добијеним у ТХМ-у, корпус питања и одговора садржи 494.288 токена, од чега 187.544 токена припада питањима, а 306.744 одговорима, док референтни нормативни корпус садржи 129.586 токена.

4.1. Дужина и структура питања и одговора

Дужина питања и одговора анализирана је као један од показатеља њихових разлика у обиму и структури. За сваку јединицу утврђен је број речи, на основу којег су израчунати основни описни статистички показатељи (Табела 1). На нивоу појединачних Q&A парова, испитан је однос дужине одговора и питања. Просечна дужина реченице израчуната је на основу броја речи и броја реченица, након аутоматске сегментације, уз претходно дефинисану листу скраћеница пронађених у корпусу код којих тачка не представља границу реченице.

Табела 1: Основни показатељи дужине питања и одговора.

Показатељ	Питања	Одговори
Број јединица	4.124	4.124
Просечан број речи	39,5	65,4
Медијана	29	47
Стандардна девијација	36	60,9
Просечна дужина реченице	24,1	28,5

Институционални одговори су у просеку 1,65 пута дужи од питања. Разлика је видљива и у медијанама, које износе 47 речи за одговоре и 29 речи за питања, док већа стандардна девијација код одговора указује на већу варијабилност њихове дужине. Просечна дужина реченице такође је већа у одговорима (28,5 речи) него у питањима (24,1 реч). Резултати показују тенденцију институције ка формулисању опширнијих одговора.

4.2. Лексичка разноврсност

Лексичка разноврсност испитана је применом библиотеке LexicalRichness (Shen, 2026) и три мере: *Type-Token Ratio (TTR)*, *Measure of Textual Lexical Diversity (MTLD)* и *Hypergeometric Distribution Diversity (HD-D)* (Табела 2)(Shen, 2022).

Табела 2: Индекси лексичке разноврсности питања и одговора.

<i>Мера</i>	<i>Питања</i>	<i>Одговори</i>
TTR	0.0769	0.0359
MTLD	97.2111	106.9765
HD-D	0.8720	0.8873

Вредности TTR-а су ниже у поткорпусу одговора, што је очекивано с обзиром на његову већу укупну дужину, будући да је ова мера осетљива на величину текста (Durán et al., 2004). За поређење поткорпуса релевантније су мере MTLD и HD-D, које показују нешто већу лексичку разноврсност у институционалним одговорима. Овај резултат не значи нужно да је њихов језик лексички богатији или мање стереотипан, већ показује да одговори садрже релативно шири распон различитих лексичких јединица. Удео лема које се у корпусу јављају само једном (hapax legomena) даје додатни увид у лексички састав поткорпуса. Он је највећи у питањима (42,78%), док је у одговорима знатно мањи (30,15%) и готово једнак уделу у референтном нормативном корпусу (30,18%). Сличност одговора и нормативних текстова, уз подударање дела најфреквентнијих лема, може указивати на ослањање институционалних одговора на језик нормативних докумената.

4.3. Карактеристичне лексеме и конструкције

Дистрибуција врста речи показала је да именице чине приближно 33% токена у одговорима, наспрам 28% у питањима, док се један глагол просечно јавља на 3,6 именица у одговорима и 2,9 именица у питањима. Ови резултати указују на већу заступљеност номиналних образаца у језику институције, што се може повезати са номиналношћу као једном од карактеристика бирократског језика. Анализом специфичности додатно се издвајају леме *акт*, *програм*, *склад*, *лице* и *члан*, које се у одговорима често јављају у типичним административним терминолошким конструкцијама, као што су **у складу са**, *правно лице* и *члан пројектног тима*, *члан Акта* (Слика 3).

Предлошки изрази такође показују изражену специфичност за институционалне одговоре, нарочито *у складу са* (200), *у погледу*, *у вези са*, *у оквиру* и *у случају*. Њихова честа употреба

[[[srlemma = "sklad"] [0.50] [srlemma = "član"] [0.50] [srlemma = "akt"] [0.50] [srlemma = "program"] [1] * @ [srlemma = "angažovanje"]]] [[srlemma = "angažovanje"]]]

Left context	Pivot	Right context
ukoliko je to u skladu sa potrebama Projekta. Važno je da	[angažovanje] planirate u skladu sa članom 7. Akta programa	PRIZMA. Ukoliko
U slučaju angažovanja fizičkih i pravnih lica kao pružaoца услуга, njihovo	[angažovanje] mora biti u skladu sa članom 19. Akta programa	. Pružaoци услуга
efektivnih meseci angažovanja na projektu koji su prikazani u Gantogramu. U	skladu sa članom 15. Akta programa PRIZMA [angažovanje]	članova tima mog
u odnosu na postavljeno pitanje. Podsećamo da na mesečnom nivou u	skladu sa članom 7. Akta programa PRIZMA nije moguće [angažovanje]	sa više od
stoji mogućnost angažovanja fizičkih i pravnih lica kao pružaoца услуга, njihovo	[angažovanje] mora biti u skladu sa članom 23. Akta programa	. Pružaoци услуга

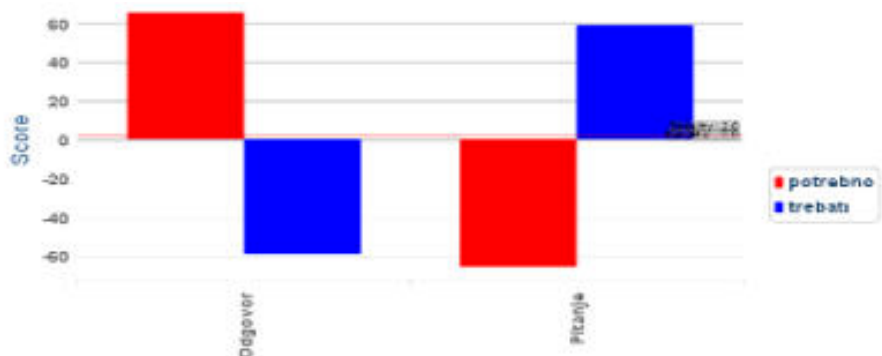
Слика 3: Примери употребе специфичних лема у институционалним одговорима (конкорданца).

чини институционални језик формалнијим. Значајан је и израз *у вези са*, који се у административним текстовима често користи као општији начин успостављања односа између појмова, што је у складу са запажањима Чудомировића (2025).

Разлике се уочавају и у употреби заменица и модалних глагола. У одговорима је изразито специфична и заменица *ви*, док се у питањима чешће јављају *ја* и *ми*. Лема *лице* има веома висок индекс специфичности у одговорима (107,3) и најчешће се користи за неутрално означавање особа, што се може повезати са деперсонализацијом карактеристичном за административни стил.

У одговорима је нешто специфичнији глагол *морати*, којим се формулишу обавезе и услови, док се у питањима истичу *моћи* и *требати*, што одражава њихову усмереност на могућности. Нарочито се издваја лема *потребно*, специфична за одговоре, док је *требати* специфичније за питања (Слика 4). Ова разлика показује да институција за обавезе и услове чешће користи безличну конструкцију *потребно је*, док апликанти исто значење изражавају непосреднијим глаголом *требати*.

Уочена је занимљива разлика у избору условних везника *ако* и *уколико*: облик *уколико* је изразито специфичан за одговоре, а *ако* је специфичнији за питања. У односу на Корпус савременог српског језика (SrpKor) (Витас et al., 2025), релативна фреквенција облика *уколико* у институционалним одговорима већа је приближно 19 пута, док је употреба облика *ако* за око 40% мања. Иако *уколико* није семантички прецизнији облик, његова честа употреба у институционалним одговорима може се повезати са стилском маркираношћу и тежњом ка формалнијем језичком изразу. Слична склоност ка дужим језичким облицима уочава се и у



Слика 4: Специфичност глагола потребно и требати у поткорпусима питања и одговора.

избору предлога *са* уместо краћег *с*, који је у бројним примерима граматички могућ.

5. Дискусија и закључна разматрања

Резултати спроведених анализа потврдили су да постоји разлика између начина на који апликанти и институција користе језик у истом комуникативном контексту. Институционални одговори су дужи и формалнији, а њихов језик чешће користи именице, укључујући стручну терминологију и устаљене административне конструкције. Питања су, с друге стране, краћа и непосреднија, и чешће су усмерена на конкретне недоумице апликаната. Ове разлике одражавају различите улоге учесника: апликанти траже појашњења правила, док институција даје њихова појашњења.

Резултати такође показују да се језик институционалних одговора у доброј мери ослања на језик нормативних докумената. Одговори и нормативни текстови деле типичне правно-административне изразе и сличан удео нискофреквентних лема. Такав начин формулисања омогућава доследно и прецизно позивање на правила јавних позива, али на тај начин истовремено може израз учинити сложенијим.

Развијени корпус пружа основу за даља истраживања институционалног дискурса и његове употребе у комуникацији са научном заједницом. Уочени језички обрасци могу бити полазиште за испитивање разумљивости институционалних одговора,

као и за будући развој алата за подршку комуникацији у системима финансирања научноистраживачких пројеката.

Захвалност

Рад представља део истраживања које се спроводи у оквиру мастер рада на мултидисциплинарним студијама при Универзитету у Београду, на студијском програму Рачунарство у друштвеним наукама. Захвалност за помоћ и подршку дугујем менторки, проф. др Ранки Станковић, редовном професору на Рударско-геолошком факултету Универзитета у Београду.

Извори података

Lucas Shen. 2022. *LexicalRichness: A small module to compute textual lexical richness*. <https://lexicalrichness.readthedocs.io/en/latest/>.

Фонд за науку РС. 2020–2025. *Документација јавних позива за финансирање научноистраживачких пројеката*. Фонд за науку Републике Србије, <https://fondzanauku.gov.rs/>.

Литература

Aryan Ahuja and Aditya Wadhwani. 2025. Bridging the Literacy Gap in Government Communication: An NLP Approach to Plain Language Translation, booktitle = Proceedings of the International Symposium on Artificial Intelligence and Natural Language Processing. In *2025 20th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, pages 1–6.

Douglas Biber and Susan Conrad. 2009. *Register, Genre and Style*. Cambridge University Press, Cambridge.

Paul Drew and John Heritage. 1992. *Talk at Work: Interaction in Institutional Settings*. Cambridge University Press, Cambridge.

Pilar Durán, David Malvern, Brian Richards, and Ngoni Chipere. 2004. *Developmental Trends in Lexical Diversity*. *Applied Linguistics*, 25(2):220–242.

- Norman Fairclough. 1989. *Language and Power*. Longman, London.
- Serge Heiden. 2010. [The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme](#). In *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation*, pages 389–398.
- Duška Klikovac. 2008. *Jezik i moć: Ogledi iz sociolingvistike i stilistike*. Biblioteka XX vek, Beograd.
- Tony McEnery and Andrew Hardie. 2012. *Corpus Linguistics: Method, Theory and Practice*. Cambridge University Press, Cambridge.
- Lucas Shen. 2026. [Shades of Neutrality: Detecting Media Slant in a Political Monopoly Using Quotation Accuracy](#). *ACM Transactions on Intelligent Systems and Technology*.
- Ranka Stanković, Branislava Šandrih, Cvetana Krstev, Miloš Utvić, and Mihailo Škorić. 2020. [Machine Learning and Deep Neural Network-Based Lemmatization and Morphosyntactic Tagging for Serbian](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3954–3962.
- Душко Витас, Ранка Станковић, and Цветана Крстев. 2025. [О фамилији корпуса савременог српског језика СрпКор](#). In *Proceedings of the International Conference South Slavic Languages in the Digital Environment JuDig: Thematic Collection of Papers*, pages 3–27, Belgrade. University of Belgrade – Faculty of Philology.
- Душка Кликовац. 2001. [О бирократизацији српског језика](#). *Наш језик*, XXXIV(1–2):85–110.
- Бранко Тошовић. 2002. *Функционални стилови*. Београдска књига, Београд.
- Јован Чудомировић. 2025. [О семантици и функционалности предлошког израза у вези \(с \(а\)\) у савременом српском језику](#). *Јужнословенски филолог*, 81(1):95–112.

Linguistic Analysis of Q&A Communication in Public Calls for Research Project Funding

Durđica Perović

Abstract

The study examines the linguistic features of institutional communication in Q&A systems accompanying public calls for research project funding. These systems represent a structured communication channel between applicants and the institution during the process of submitting research project proposals.

The study is based on a corpus compiled from publicly available documentation of the Science Fund of the Republic of Serbia, comprising 4,124 pairs of researchers' questions and corresponding institutional answers, with a total of 494,288 tokens, collected across 9 public calls. Using methods from corpus linguistics, textometry, and Natural Language Processing (NLP), the study examines differences between questions and answers in terms of their length and structure, lexical diversity, grammatical and syntactic features, and the use of characteristic lexical items and constructions. Particular attention is paid to linguistic features associated with bureaucratic language and the administrative functional style, including nominalisation, characteristic prepositional constructions, and modal constructions.

The results reveal clear differences between the language of researchers' questions and institutional answers. Selected findings are further examined through comparison with data from the reference corpus of the Serbian language (SrpKor) and with a corpus compiled from publicly available normative documents of the Science Fund of the Republic of Serbia that define the requirements and procedures for applying to the respective public calls. The application of computational methods to the analysis of this structured corpus enables a more precise understanding of the specific linguistic patterns that characterise institutional discourse in contrast to applicants' questions. The findings may contribute to a better understanding of official instructions and provide a basis for improving institutional communication.

Keywords: institutional discourse, bureaucratic language, corpus linguistics, textometry, natural language processing, research funding

Works are licensed under the terms and conditions of the Creative Commons Attribution 4.0 International License.

<https://creativecommons.org/licenses/by/4.0/>



Кориснички интерфејс за унос ономастичких података: искуства, изазови и унапређење модела

Снежана Петровић¹[0000-0003-1549-7294], Владимир
Поломац²[0000-0001-5140-9330], Ана Шпановић¹[0000-0001-7368-9660], Матија
Нешовић¹[0009-0004-5822-6675], Јована Тодорић¹[0009-0002-0706-6312], Ленка
Бајчетић¹[0000-0002-2005-5426]

¹Српска академија наука и уметности, Институт за српски језик, Србија,
{snezzanaa|tesicana|nesovic1998|jovanatodoric080|lenka.bajcetic}@gmail.com

²Универзитет у Крагујевцу, Филолошко-уметнички факултет, Србија,
vladimir.polomac@gmail.com

Апстракт

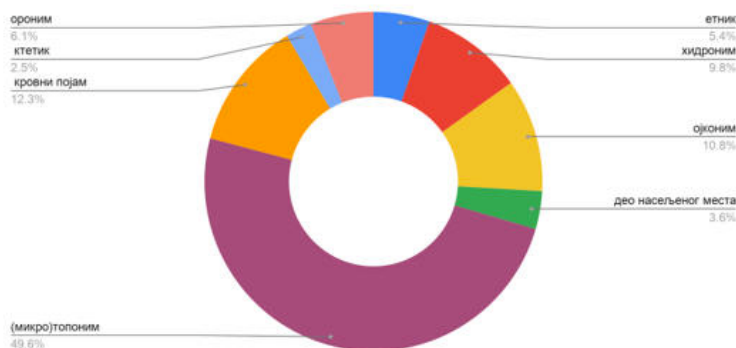
Рад приказује даље кораке у развоју корисничког интерфејса за унос ономастичких података у оквиру пројекта дигитализације ономастичке грађе Одбора за ономастику САНУ, чији је циљ дигитализација око 750.000 података прикупљених теренским истраживањем током више деценија. У првој фази уноса обрађено је 11.464 топонима различитих категорија. Интерфејс, развијен у окружењу Python/Django, унапређен је низом измена: поље тип сада омогућава вишеструки избор, додата су поља за број кућа, број глава и предлошко-падежну конструкцију, уведено је поље година за хронолошко позиционирање грађе, као и функција спајања повезаних листића. Модел сада подржава истовремени рад више корисника уз систем персоналних сесија, спречавање дуплирања уноса и маркирање података за ревизију. Унете су прецизне геолокације за 652 насеља са Косова и Метохије, што омогућава визуелизацију и просторну анализу грађе на интерактивној мапи. Рад разматра и изазове у одређивању стандардизованих облика топонима и разграничење граничних категорија, на пример оронима и хидронима, као и проблеме одређивања рода и броја код етимолошки непрозирних лексема. На крају су представљени планови за развој бета верзије јавно доступне платформе за отворени приступ дигитализованој грађи.

Кључне речи: дигитализација, ономастика, кориснички интерфејс, топоними, геолокација, дијалекатска грађа, база података.

1. Проширени апстракт

У овом раду ће бити приказани даљи кораци у развоју и унапређивању корисничког интерфејса за унос ономастичких података у оквиру пројекта дигитализације ономастичке грађе Одбора за ономастику САНУ. Пројекат је настао са основним циљем да теренску грађу сакупљању током више деценија, која броји око 750 000 ономастичких података, дигитализује и омогући њихово повезивање и приказивање уз помоћ различитих врста претрага, као и кроз визуелизацију података у виду дигиталних мапа. О другим типовима дијалекатске грађе и могућој (ретро)дигитализацији в. (Петровић *et al.*, 2025b), а о сличним пројектима дигитализације ономастичког материјала в. (Waldispühl *et al.*, 2020; Миленковска *et al.*, 2025).

У досадашњој, првој фази дигитализације акценат је био стављен на унос и обраду топонима. До писања овог апстракта унето је 11464 топонима различитих категорија, међу којима су: микротопоними (5682), ојконими (1235), кровни појмови (1414), хидроними (1119), ороними (694), етници (616), ктетици (289) и делови насељеног места (415). Процентуална дистрибуција категорија топонима приказана је на графикону (Слика 1). Унос антропонима планиран је за наредну фазу рада.



Слика 1: Процентуална дистрибуција категорија топонима.

Модел корисничког интерфејса за унос ономастичких података имплементиран је помоћу програмског језика Python и библиотеке Django и омогућава вишекорисничко окружење за унос, преглед и кориговање ономастичких података, као и података везаних за локације (географске координате, измењени називи ме-

ста и општина и др). Даљи развој интерфејса одвија се по принципу итеративног унапређивања према коме свака нова фаза уноса ономастичких података утиче на могуће измене у моделу. Итеративно унапређивање модела било је неопходно будући да је ономастичка грађа која се дигитализује изразито хетерогена и да је сакупљана од стране већег броја записивача у различитим временским периодима. Разноликост записа била је условљена и ограниченошћу самих упутстава за сакупљање грађе која нису могла унапред предвидети сваки појединачни случај на терену, те су записивачи неретко бележили додатне податке, изостављене из упутстава, које су сматрали важним за дату локацију. Стога ни анализа узорка грађе на основу које је урађена прва верзија интерфејса (в. (Петровић et al., 2025a)) није могла обухватити све њене специфичности, већ су се потребе за изменама, на основу којих се модел циљано унапређивао, уочавале тек кроз непосредан рад са ономастичким подацима.

Претходни модел обухватао је пре свега основна поља за унос ономастичке грађе. На десној страни екрана налазе се поља за унос следећих података: акцентовани облик насловне речи, стандардизовани облик насловне речи, опис ономастичког податка, коси падеж, кровни појам и коментар. Поља *род*, *број* и *тип* попуњавају се одабиром одговарајућих опција из падајућег менија. Поље *алтернативни називи* резервисано је за унос фонолошких, морфолошких и, евентуално, лексичких варијанти одређеног ономастичког податка забележеног на терену, док поље *упућенице* омогућава једноставније повезивање реферишућих листића. Поред ових поља, издвајају се још два истог назива: *потребна ревизија*. Овако названо поље смештено у горњем левом делу екрана омогућава да се назначи да ли су потребне исправке података који су везани за одговарајући скенирани листић. За разлику од њега, поље смештено у доњем десном углу указује на потребу да се провере и/или измене унети ономастички подаци са скенираних листића (Слика 2).

На основу анализе до сада унетих података, дефинисане су следеће измене за нову верзију корисничког интерфејса.

Категорија *тип* ономастичког податка, која је у почетној верзији била постављена као јединична вредност, редефинисана је тако да сада омогућава истовремени избор више опција из унапред дефинисаног скупа вредности. Ова измена показала се не-

The screenshot displays the JUDIG web application interface. On the left, there is a preview of a handwritten note on a grid background. The text in the note is: "Турћеваци", "назив становника села", "Турћеваци", "115", and "с. Турћеваци". Below the note is the file name "008_14_0021_000587_04_00001.png". On the right, there is a data entry form with the following fields: "Адресовано:" (Turčevanci), "Стандардно:" (Turčevanci), "Опис:" (назив становника села Турћеваци), "Код наред:" (-4), "Кривој појам:" (1), "Коментари:" (empty), "Упућеница:" (empty), "Алтернативни називи:" (empty), and "Ред:" (empty). At the bottom right, there are buttons for "Сачувај" and "Потребна ревизија".

Слика 2: Претходна верзија модела корисничког интерфејса.

опходном у случајевима када исти податак покрива више ономастичких категорија (Слика 3).

The screenshot displays the JUDIG web application interface. On the left, there is a preview of a handwritten note on a grid background. The text in the note is: "Фабја репа", "песачи, ораница", "и део села", "114", "М.А.", and "с. Врхуница". Below the note is the file name "008_14_0021_000021_04_00004.png". On the right, there is a data entry form with the following fields: "Адресовано:" (Фабја репа), "Стандардно:" (Фабја репа), "Опис:" (песачи, ораница и део села), "Код наред:" (empty), "Кривој појам:" (empty), "Коментари:" (empty), "Упућеница:" (empty), "Алтернативни називи:" (empty), "Тип:" (empty), "Ред:" (empty), "Бр. кућа:" (empty), "Бр. глава:" (empty), "Ревизија:" (empty), "Спој са претходним:" (empty), and "Датум:" (empty). At the bottom right, there are buttons for "Сачувај" and "Потребна ревизија".

Слика 3: Хидроним, део насељеног места и (микро)топоним.

Поред тога, модел је проширен и новим пољима. Код обележавања ономастичких података додата су поља *број кућа*, *број глава* и *предлошко-падежна конструкција*. Поље *година* додато је као нова опција при обележавању ономастичких листића и оно сада омогућава прецизно хронолошко позиционирање ономастичке грађе (Слика 4).

Уведена је и функција *спој са претходним податком*, неопходна стога што се њоме повезују различити листићи који се односе на један исти податак. То је посебно релевантно у случа-

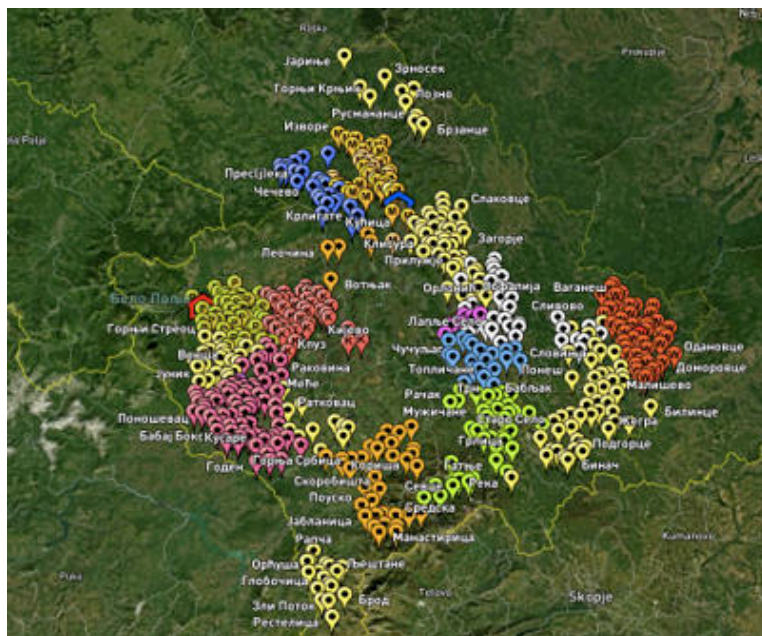
корисник отворио или прегледао остају резервисани за њега током трајања сесије. На техничком плану, уведене су персоналне сесије које бележе не само које је листиће корисник обрадио већ и које је прегледао, омогућајући тако слободно кретање уназад кроз сопствену историју рада. Посебна пажња посвећена је томе да поновни преглед већ обрађеног листића не доведе до губитка унетих података, уз истовремено очување могућности измене података уколико је то потребно. Допуњен је и систем ревизије: ономастички листићи али и појединачни подаци могу бити обележени маркером *потребна ревизија*, чиме се означава које податке треба додатно проверити. Овај механизам показао се нарочито корисним у раду са сложенијом ономастичком грађом.

Модел интерфејса унапређен је на плану геолоцирања и визуелизације података. Унете су прецизне координате за 652 локације са подручја Косова и Метохије, укључујући и она насељена места која су у међувремену променила назив или која данас више не постоје. На тај начин је омогућено директно приказивање података на интерактивној мапи, њихово прегледање у свакој локацији понаособ, као и просторна анализа ономастичке грађе (Слика 6). Детаљније о ономастичкој грађи са простора Косова и Метохије в. (Петровић-Савић, 2023).

Иако су изазови који се јављају при унапређивању модела корисничког интерфејса централна тема рада, осврнућемо се и на неке недоумице које су се јавиле током уношења ономастичких података, а које се тичу њиховог прецизног обележавања и дефинисања.

Међу њима се издвајају топоними на албанском језику и топоними са варијантама на другим језицима, за које је било потребно одредити одговарајући начин бележења и стандардизације унутар постојећег система поља.

У појединим се случајевима поставља питање одређивања типа ономастичког податка. На пример, треба прецизирати какви се све облици рељефа могу означити као ороними – да ли само типична узвишења (попут брда и планина) или и другачије конфигурације терена (попут јаруга, долина, удолина, клисура, стена, пећина, кршева, камењара)? Слично томе, није најјасније да ли је називе различитих објеката за водоснабдевање (као што су чесме или бунари) упутно сврстати у микротопониме или хидрониме. Како се ови термини у литератури дефинишу на ра-



Слика 6: Визуелизација и геолоцирање података.

зличите начине, потребно је одабрати једно решење и доследно га спровести кроз читаву грађу.

Понекад није једноставно предложити стандардизовани облик неког назива. Дијалекатски облици се неретко веома разликују од стандардних, нарочито на фонолошком плану, где се могу приметити различите гласовне алтернације (попут умекшавања, метатезе, обезвучавања, губљења гласова и сл.), као и фонеме које стандардни српски језик не познаје (нпр. полугласник у призренско-тимочким говорима). У таквим случајевима обрађивач треба да процени када је оправдано успостављати (стварни или хипотетички) стандардни облик назива, посебно водећи рачуна о томе да ће базу претраживати и корисници којима ће стандардни облици бити познатији него дијалекатски. На пример, називи *Лојзе*, *Л'ескопат*, *Рунаће*, *Къмберова чешма* и *Петора* могу се стандардизовати као *Лозје*, *Лескопад*, *Рупаче*, *Камберова чесма* и *Петорах* (у опису овог податка је назначено да је локалитет добио име по пет ораха). Да би се стандардизација топонима доследно спровела, а тиме и олакшало претраживање базе, у неким случајевима се мора променити и граматички род именице, као и придева, ако уз њу стоји (нпр. име сеоске махале *Царлетово*

ма'ало у стандардизованој форми би гласило *Царлетова махала*). О одређеним проблемима и решењима формирања стандардизованих облика у дијалекатској грађи в. (Петровић and Тасовац, 2016).

Појединим називима се не могу са сигурношћу приписати одговарајуће вредности граматичких категорија рода и броја. То је обично случај са етимолошки непрозирним лексемама или познатим речима које су у неком говору морфолошки другачије уобличене у односу на стандард. Овакве недоумице често се односе на именице које се у номинативу множине мушког рода или у номинативу једине средњег рода завршавају вокалом -е (*Долове, Жирове*). Наилази се и на необичне примере синтагматског типа где придев не конгруира са именицом уз коју стоји, па је одређивање морфолошких вредности веома отежано, нпр. *Јапетове путинчича*.

Коначно, у завршном делу рада биће представљени планови за бета верзију јавно доступне платформе која ће омогућити отворени приступ дигитализованој ономастичкој грађи. Платформа је тренутно у фази израде, а њен циљ је да истраживачима и другим заинтересованим корисницима обезбеди претраживо и визуелно квалитетно окружење за преглед грађе Одбора за ономастику САНУ.

Захвалница

Овај рад финансирало је Министарство науке, технолошког развоја и иновација Републике Србије према Уговору број 451-03-33/2026-03/200174, који је склопљен са Институтом за српски језик САНУ.

References

Michelle Waldispühl, Dana Dannells, and Lars Borin. 2020. *Material Philology Meets Digital Onomastic Lexicography: The Nordicon Database of Medieval Nordic Personal Names in Continental Sources*. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 860–867, Marseille, France. European Language Resources Association.

- Соња Миленковска, Ангелина Панчевска, and Давор Јанкулоски. 2025. Дигитална обработка на македонскиот дијалектен материјал во Истражувачкиот центар за ареална лингвистика „Божидар Видоески“ при МАНУ. In *Slovanské dialekty v době digitálních technologií: Nářeční prameny a jejich současné zpracování = Slavic Dialects in the Age of Digital Technologies: Dialect Sources and Their Current Processing*, pages 273–282, Праг, Чешка. Slovanský ústav AV ČR.
- Снежана Петровић, Мирјана Петровић-Савић, Ана Шпановић, Ленка Бајчетић, Матија Нешовић, and Јована Тодорић. 2025a. [Дигитализација грађе Одбора за ономастику САНУ – значај, циљеви и први кораки](#). In *JUDIG – Proceedings*, pages 337–348, Београд, Србија. Универзитет у Београду – Филолошки факултет.
- Снежана Петровић and Тома Тасовац. 2016. [Лексикографска анотација нестандардних облика у служби претраживости електронских речника](#). In *Лексикологија и лексикографија у светлу савремених приступа*, pages 155–168. Институт за српски језик САНУ, Београд, Србија.
- Снежана Петровић, Ана Шпановић, Мирјана Петровић-Савић, and Ленка Бајчетић. 2025b. Serbian dialects in the digital age: Present and future practices. In *Slovanské dialekty v době digitálních technologií: Nářeční prameny a jejich současné zpracování = Slavic Dialects in the Age of Digital Technologies: Dialect Sources and Their Current Processing*, pages 283–294, Праг, Чешка. Slovanský ústav AV ČR.
- Мирјана Петровић-Савић. 2023. [Косовско-метохијска грађа при Одбору за ономастику САНУ: преглед обрађених и преосталих пунктова, послови скенирања, дигиталног уноса и интернет презентације](#). In *Заштита, очување и афирмација српског културног наслеђа на Косову и Метохији II*, pages 603–624. Српска академија наука и уметности, Београд.

The User Interface for Onomastic Data Entry: Experiences, Challenges, and Model Improvements

Snežana Petrović, Vladimir Polomac, Ana Španović, Matija Nešović, Jovana Todoric, Lenka Bajčetić

Summary

This paper presents further developments in the interface for entering onomastic data within the Committee for Onomastics of the Serbian Academy of Sciences and Arts' digitization project, which aims to digitize roughly 750,000 items of field-collected material. In the first phase, focused on toponyms, 11,464 entries have been processed to date. Built in Python/Django, the interface now allows multiple simultaneous values for the type field, adds fields for number of inhabitants, number of households, and prepositional phrases, introduces a year field for chronological positioning, and supports merging related index cards. It also enables real-time multi-user work, session-based tracking, and revision flagging. Precise geolocation has been added for 652 locations in Kosovo and Metohija, enabling map-based visualization. The paper further discusses classification and standardization challenges in dialectal material, and outlines plans for a beta version of a public platform.

Keywords: digitization, onomastics, user interface, toponyms, geolocation, dialectal material, database

Works are licensed under the terms and conditions of the Creative Commons Attribution 4.0 International License.

<https://creativecommons.org/licenses/by/4.0/>



Мерење квалитета аутоматског пресловљавања латинице на ћирилицу

Вељко Миљанић¹, Јован Турањанин¹,
Михаило Шкорић¹[0000-0003-4811-8692]

¹ Ћирилица Лаб, Србија, {veljkomilj|jovant|procesaur}@gmail.com

Апстракт

Аутоматско пресловљавање латинице на ћирилицу је важан задатак у оквиру развоја језичких технологија за српски језик и има широку практичну примену. Овај проблем није тривијалан, пре свега због изузетака за стране речи, мерне јединице и сл. као и због вишеструкости латиничног писма и то нарочито у његовом крњем облику који користи само ASCII знакове (без дијакритика). Недостатак досадашњег развоја алата за пресловљавање је одсуство стандардног тест сета за анализу тачности алата. Из овог разлога је тачност постојећих алата непозната. У овом раду представљамо јавно доступан тест сет и алате за мерење квалитета транслитерације латинице у ћирилицу. Такође, приказаћемо и анализу квалитета одабраних доступних алата за пресловљавање отвореног кода доступног на вебу. Резултати показују да већина алата има тривијалну имплементацију која неправилно пресловљава изузетке као и да не постоји алат који са задовољавајућом тачношћу може да преслови ASCII (крњу) латиницу која је веома заступљена на друштвеним мрежама. Управо из тог разлога, овај рад поставља основу за прецизну евалуацију алата за пресловљавање и развој робуснијих решења у дигиталној обради српског језика.

Кључне речи: ћирилица, пресловљавање, алати за обраду језика

1. Увод

Проблем пресловљавања са латинице на ћирилицу има дугу традицију (Aurousseau, 1953; Gerych, 1965), и од великог је значаја у српској писмености и култури (Bugarski, 2021). Ћирилица се често налази у мањинском положају у дигиталном простору, где доминира латиница (Salamurović, 2019; Tvrdišić, 2023). Због тога је развој поузданих алата за аутоматско пресловљавање од

суштинског значаја за очување и равноправну употребу ћирилице у савременим технологијама.

Пресловљавање у ћирилицу није тривијалан задатак због вишезначности латиничног писма (нпр. Танјуг, надживети, инјекција), изузетака од пресловљавања (мерне јединице, римски бројеви и сл.) као и употребе изворног записа страних имена ентитета и стручних термина (нпр. *Spase-X*). Нарочито тежак проблем представља и широка употреба ASCII латинице на друштвеним мрежама и у другој неформалној комуникацији.

Главни недостатак досадашњег рада на алатима за пресловљавање је одсуство стандардног и јавно доступног тест сета и пратећих алата за мерење тачности. Иако је до сада направљено на десетине алата, њихова тачност је остала непозната, а сам процес њиховог развоја методолошки неуједначен. Уместо примене ригорозних научних процедура, које подразумевају коришћење стандардних скупова података и прецизних метрика, развој се до сада углавном ослањао на појединачне, несистематичне покушаје.

У наставку најпре представљамо алате који су били предмет ове евалуације, уз образложење критеријума за њихов одабир. Потом описујемо скупове података коришћене за тестирање и објашњавамо њихове специфичне карактеристике и припрему. Следи део посвећен евалуацији и резултатима, где приказујемо методологију мерења, добијене резултате и анализу најчешћих грешака. На крају, у закључку, дајемо процену тренутног стања у области аутоматског пресловљавања, разматрамо потребу за новим и бољим алатима и предлажемо могуће правце даљег развоја.

2. Алати

За евалуацију су одабрани различити алати за пресловљавање са латинице на ћирилицу. Као бејслајн узето је директно пресловљавање слова у слово, без додатних правила, какво обезбеђују пакети *srbai* и *srtools*. Ово представља најједноставнији приступ и служи као полазна тачка за поређење. Критеријум за одабир осталих алата био је практична употребљивост: алати који су јавно доступни и који се лако могу интегрисати у аутоматизовани ток рада, као и они који нуде више од једноставног мапирања слова у слово (бејслајн). На овај начин обезбеђено је да евалуација обухвати и најједноставније и напредније приступе,

што омогућава свеобухватну процену тренутног стања у области пресловљавања. Тестирани алати побројани су Табели 1.

Табела 1: Списак евалуираних алата са изворима и технологије у оквиру којих су настали.

РБ	Назив алата	Извор	Технологија
1	Serbian-AI-Society/SrbAI	GitHub	Python
2	andrejr/srtools	GitLab	
3	eevan78/translit	GitHub	GO
4	turanjanin/cirilizator	GitHub	JavaScript
5	ArtBIT/youconv	GitHub	
6	exvorn/serbian-transliteration	GitHub	
7	ivebe/CyrLatConverter	GitHub	
8	pioniredu/preslovljivac-js	GitHub	
9	raleksandar/pravopis	GitHub	PHP
10	turanjanin/serbian-transliterator	GitHub	
11	turanjanin/serbian-language-tools	GitHub	RUST
12	filiparag/translitRS	GitHub	

3. Подаци

За евалуацију успешности алата коришћена су четири скупа текстова, пажљиво одабрана због свог квалитета и релевантности. Скупови су побројани и описани у Табели 2.

Табела 2: Називи и описи прикупљених скупова са постотком ћирилице и величинама развојних и тест скупова у милионима речи.

Назив	Опис	Ћир.	Разв.	Тест
Новине	Новински чланци објављени на веб порталу <i>Полиџике</i> , у периоду јануар–април 2025. год.	99,8	0,46	0,71
Википедија	Српска <i>Википедија</i> , одабрани чланци, пресек 1. мај 2025. год.	95,4	1,76	1,81
Знање	Докторске дисертације са платформе НаРДуС, објављене у оквиру корпуса <i>Знање</i>	95,6	3,22	3,66
Редит	Објаве користика са друштвене мреже <i>Редит</i> на ћирилици	99	0,23	0,35

Сви скупови су оригинално писани на ћирилици, што је био примарни критеријум за избор. Поред тога, текстови из ових извора прошли су процес лектуре или рецензије, па се може очекивати висок степен језичке исправности и стандарда. На тај начин обезбеђено је да тестирање буде засновано на поузданим и квалитетним подацима, без утицаја шума или неконзистентности у изворном материјалу.

Чланци из онлајн издања дневних новина *Политика* прикупљени су веб-скрејпингом са њиховог портала, у периоду од јануара до априла 2025. године. Овај скуп је одабран јер представља актуелне текстове високог језичког квалитета, који су прошли редакцијску обраду и лектуру. На тај начин обезбеђен је корпус који одражава савремени језички стандард, са богатим стилским и тематским варијацијама.

Други скуп података чини ексцерпт најновијег извоза Викимедијине базе за српску Википедију од 1. маја 2026. године (Škorić and Palma, 2026). Одабрани су чланци минималне дужине од 50 речи, при чему је постотак текста на ћирилици морао бити најмање 85%. Ови чланци су изабрани јер су оригинално писани на ћирилици, али садрже и елементе који се не пресловљавају (јединице мере, римски бројеви, латински и други страни изрази), што представља додатни изазов за евалуацију алата. Трећи скуп обухвата докторске дисертације писане на српском језику (и ћириличној писму) са платформе НаРДуС, преузете из корпуса *Знање* (Škorić and Janković, 2024) и филтриране по истом критеријуму као и други скуп (постотак ћириличних речи од 85%). Четврти скуп је настао као ексцерпт коментара на српским подредитима (*r/serbia*, *r/srbija* и *r/serbiancringe*). Подаци су прикупљени из обраде редовог извоза (Hoang, 2024), при чему су филтрирани коментари са постотком ћириличних речи већим од 85%.

Други, трећи и четврти скуп заједно садрже око 220 милиона речи и објављена су на платформи *HuggingFace*¹. Ресурс је подељен на тренинг (95%), развојни (2%) и тест скуп (3%). Тренинг и развојни скупови намењени су потенцијалном обучавању модела за пресловљавање, док је тест намењен искључиво за евалуацију система за пресловљавање, укључујући и експеримент представљен у овом раду.

¹ <https://huggingface.co/datasets/procesaur/cirilica>

4. Евалуација и резултати

Први корак у евалуацији био је да се прикупљени ћирилични текстови једнозначно преслове на латиницу, карактер по карактер. Овај поступак не подразумева губитак информација нити амбигуитет, јер сваки знак има јасну латиничну мапу. На тај начин добијен је контролисани корпус латиничних текстова, који је потом коришћен као улаз за различите алате за пресловљавање. Сваки алат је враћао текст у ћирилицу, а добијени резултати су упоређивани са оригиналним ћириличним текстом.

Успешност је мерена пре све свега стопом грешке у речима (енг: Word Error rate, WER). Резултати су израчунати тако што су очекивана секвенца и транслитерирана секвенца токена поравнате Питон библиотekom Левенштајн (Levenshtein) а затим је стопа грешке израчуната стандардном формулом:

$$WER = (S + D + I)/N \quad (1)$$

Где је S (Substitutions) број погрешних замена токена, D (Deletions) број обрисаних токена, I (Insertions) број уметнутих токена.

Поред основног задатка, експеримент је проширен тежим сценаријем: пресловљавање *ошмишаної* латиничног текста (без дијакритика), који је чест у веб окружењу. Овај задатак је значајно сложенији јер захтева од алата да препозна контекст и исправно реконструише ћириличне знакове. Алати и скупови су интегрисани кроз јединствени процес заснован на *bash* скриптама и *Python* алатима, што је омогућило систематично тестирање, репродуктивност резултата и лако поређење различитих приступа.

Резултати приказани у Табели 3 показују да већина тестираних алата остварује релативно сличну просечну стопу грешке (између 2 и 3%), као и да резултати на развојном и тест скупу имају малу стопу варијације. Такође, већина алата надмашује бејслајн (мапирање карактера у карактер).

Најбоље резултате на већини тестова (са изузетком тест подскупа *Знање*), као и у просеку постигао је алат *turanjanin/cirilizator*. Овај резултат није изненађење јер се овај алат често помиње у документацијама других тестираних алата као референца и као извор одакле су преузета правила и изузеци за приликом развоја алгорита за пресловљавање.

Табела 3: Резултати алата за пресловљавање (WER) за сва четири скупа и у просеку за развојни скуп (горња половина) и тест скуп (доња половина), са просечном стопом грешке за сваки подскуп (укошено).

Назив алата	Новине	Вики.	Знање	Редит	просек
Serbian-AI-Society/SrbAI	0,30	4,64	12,92	0,66	4,63
andrejr/srtools	0,30	4,64	5,11	0,66	2,68
eevan78/translit	0,83	3,50	4,38	1,19	2,48
turanjanin/cirilizator	0,14	3,16	4,10	0,62	2,01
ArtBIT/youconv	0,16	4,34	4,88	0,64	2,51
exvorn/serbian-transliteration	0,14	4,06	4,70	0,58	2,37
ivebe/CyrLatConverter	0,17	4,66	5,12	0,68	2,66
pioniredu/preslovljivac-js	0,22	4,84	5,15	0,68	2,72
raleksandar/pravopis	5,78	10,51	12,5	4,95	8,44
turanjanin/serbian-transliterator	0,16	3,45	4,27	0,66	2,14
turanjanin/serbian-language-tools	0,18	3,97	7,16	0,73	3,01
filiparag/translitRS	0,15	3,90	6,45	0,76	2,82
<i>просек свих алата</i>	<i>0,71</i>	<i>4,64</i>	<i>6,40</i>	<i>1,07</i>	<i>3,20</i>
Serbian-AI-Society/SrbAI	0,33	4,50	8,27	0,69	3,45
andrejr/srtools	0,33	4,50	4,34	0,65	2,46
eevan78/translit	0,93	3,48	3,69	1,19	2,32
turanjanin/cirilizator	0,17	3,06	3,28	0,61	1,78
ArtBIT/youconv	0,19	4,18	4,13	0,62	2,28
exvorn/serbian-transliteration	0,17	3,92	3,89	0,58	2,14
ivebe/CyrLatConverter	0,19	4,53	4,35	0,67	2,44
pioniredu/preslovljivac-js	0,24	4,70	4,40	0,67	2,50
raleksandar/pravopis	5,82	10,43	12,05	4,96	8,32
turanjanin/serbian-transliterator	0,19	3,31	3,51	0,63	1,91
turanjanin/serbian-language-tools	0,22	3,88	6,88	0,73	2,93
filiparag/translitRS	0,19	3,82	5,86	0,76	2,66
<i>просек свих алата</i>	<i>0,75</i>	<i>4,53</i>	<i>5,39</i>	<i>1,06</i>	<i>2,93</i>

Просечна грешка по доменима (укошено, Табела 3) показује да је најлакши задатак пресловљавање новинских чланака (WER 0,75 на тест скупу), док су најзахтевнији за обраду корпуси Знање и Википедија (WER 5,39 и 4,53 на тест скупу). Разлог за овакве резултате лежи у чињеници да највећи удео у грешкама чине

стручни термини, цитати на страним језицима, мерне јединице и римски бројеви, који су високо заступљени управо у корпусу знања.

Ове налазе потврђује и квалитативна анализа 100 најчешћих грешака, приказана у Табели 4. У свим анализираним случајевима, проблем је повезан или са неадекватном детекцијом текста који је требало изоставити из процеса пресловљавања, или са шумом у тест скупу (латиничне речи или карактери уметнути у ћирилични текст).

Табела 4: Анализа 100 најчешћих грешака за алат *turanjanin/cirilizator* на подскупу *Знање*.

Тип грешке	Пример (Очекивано -> Пресловљено)	Удео
Страни изрази и цитати	Journal -> Јоурнал	51%
Оригинални текст на латиници или латинична слова у ћириличним речима (шум у корпусу)	Dobijeni rezultati pokazuju -> Добијени реултати показују	40%
Пресловљене мерне јединице	200-300 m -> 200-300 м	4%
Пресловљени римски бројеви	I -> И	4%
Није пресловљено грешком (пропуштена конверзија)	СС формација -> SS формација	1%

Највећи удео (51%) чине стране речи које су погрешно пресловљене на ћирилицу, док по 4% грешака отпада на неадекватну конверзију римских бројева и мерних јединица.

Велики део грешака (40%) односи се на речи на српском језику које су у оригиналним документима погрешно записане (уметнута латинична слова или целе латиничне речи, реченице (често цитати) у иначе ћирилични текст, или грешке у куцању које су се негативно одразиле препознавање изузетака). Овај тип грешке представља системски шум (недостатак) у самом корпусу, а не стварни функционални пропуст у раду тестираних пресловљивача.

Додатне грешке чине мерне јединице (4%) и римски бројеви (4%), који не треба да се пресловљавају на ћирилицу, али овде су грешком пресловљени.

Коначно, речи које су грешком задржане у латиничном облику (пропуштена транслитерација) чине 1% од укупног броја. Претпоставља се да је ова појава узрокована лажним позитивима унутар алгоритама за детекцију мерних јединица, римских бројева и страних речи, који су нека од слова регуларног текста грешком оставили нетакнутим.

За анализу ефикасности пресловљавања ошишане латинице коришћен је подскуп из домена вести, будући да овај корпус садржи највећи удео изворно ћириличних текстова. Сходно претходној анализи, ово гарантује да остали извори грешке, који нису везани за ошишању латиницу сведени на минимум.

Мерење је извршено искључиво над алатом *turanjanin/serbian-language-tools* који је једини пресловљивач са подршком за рестаурацију дијакритика. Резултати овог експеримента су дати у Табели 5. Пресловљавање текста без претходног враћања дијакритика даје изузетно високу стопу грешке од 15.37%, што овакав приступ чини практично неупотребљивим у реалним применама. Са друге стране, претходно враћање дијакритика смањује стопу грешке на само 1.08% што представља значајно побољшање квалитета, и намеће се као неопходност у оваквој поставци.

Табела 5: Резултати на подскупу Вести (тест) за алат *turanjanin/serbian-language-tools* са и без враћања дијакритика

Приступ	Латиница (WER)	ASCII (WER)
Пресловљавање	0,19%	15,37%
Враћање дијакритика + пресловљавање	0,40%	1,08%

Значајно је приметити и да разлика у стопи грешке између пресловљавања исправног латиничног текста (0.19%) и пресловљавања ошишане латинице уз враћање дијакритика (1.08%) потиче од несавршености у процесу рестаурације дијакритика, што отвара додатне теме за будућа истраживања.

5. Закључак

Тренутно стање у области аутоматског пресловљавања показује да, упркос великом броју доступних алата, кључни изазови и даље нису задовољавајуће решени.

Тачност при конверзији ошишане латинице није на задовољавајућем нивоу и захтева анализу ширег контекста ради успешног разрешавања вишезначности речи. Такође, ниједан тренутно доступан алат не поседује адекватну подршку за детекцију страних израза и цитата које је потребно оставити у изворном облику.

Препознавање једноставнијих нетекстуалних елемената, попут мерних јединица и римских бројева, није савршено и често доводи до лажно негативних, а понекад и до лажно позитивних резултата. Са друге стране, детекција комплекснијих нетекстуалних структура, као што су математичке формуле или серијске ознаке производа, тренутно уопште није подржана.

Поред наведених недостатака у квалитету, подједнако велики изазов представља и изражена софтверска фрагментираност. У пракси не постоји јединствено решење (универзална библиотека) које би изворно функционисало на свим популарним програмским језицима и платформама. Такође, не постоји ни јединствени референтни списак правила и изузетака, што би омогућило лакшу интеграцију и већу конзистентност алата. Недостатак таквог заједничког технолошког језгра спречава компаније да успоставе индустријски стандард који би се једноставно интегрисао при развоју и локализацији софтвера.

Ресурси

- Anh Chang Hoang. 2024. Reddit Pushshift Dataset Cleaned. https://huggingface.co/datasets/anhchanghoangsg/reddit_pushshift_dataset_cleaned. Accessed: 2026-05-27.
- Mihailo Škorić and Cosimo Palma. 2026. WikiKorpus. <https://huggingface.co/datasets/procesaur/WikiWiki>. License: CC-BY-SA-4.0.

Литература

- M. Aurousseau. 1953. Transliteration of Cyrillic script. *Nature*, 171(4360):940–940.
- Ranko Bugarski. 2021. A war of letters: The case of Serbian. *Serbian Studies: Journal of the North American Society for Serbian Studies*, 32(1):175–192.
- Iurii Gerych. 1965. *Transliteration of Cyrillic alphabets*. Ph.D. thesis, University of Ottawa (Canada).
- Aleksandra Salamurović. 2019. Тирилица из нот дед: Cyrillic script from a sociolinguistic perspective in Macedonia, Montenegro and Serbia. *Digitale Bibliothek üringen*.
- Mihailo Škorić and Nikola Janković. 2024. New Textual Corpora for Serbian Language Modeling. *Infotheca*, 24(1):71–96.
- Sara Tvrdišić. 2023. Cyrillic Heritage in the Shadows: Exploring the Diminishing Use of Serbian Cyrillic Online in the Age of AI. In *Artificial Intelligence Embraced: the future of the cultural and creative sector.*, pages 5–19.

Measuring the Quality of Automatic Transliteration from Latin to Cyrillic

Veljko Miljanić, Jovan Turanjanin, Mihailo Škorić

Summary

Automatic transliteration of Latin into Cyrillic is an important task in the development of language technologies for the Serbian language and has wide practical application. This problem is not trivial, primarily due to exceptions for foreign words, units of measurement, etc., as well as due to the ambiguity of the Latin script, especially in the form that uses only ASCII characters (without diacritics). A shortcoming of the development of transliteration tools to date is the absence of a standardised test set for analyzing the accuracy of the tools. For this reason, the accuracy of existing tools is unknown. In this paper, we present a publicly available test set and tools for measuring the quality of transliteration of Latin into Cyrillic. We will also present an analysis of the quality of selected available open source transliteration tools available on the web. The results show that most tools have a trivial implementation that incorrectly transliterates exceptions, and that there is no tool that can with satisfactory accuracy transliterate ASCII Latin, which is widely used on social networks. For this reason, this paper sets the basis for a precise evaluation of transliteration tools and the development of more robust solutions in the digital processing of the Serbian language.

Keywords: Cyrillic, transliteration, language processing tools

Works are licensed under the terms and conditions of the Creative Commons Attribution 4.0 International License.

<https://creativecommons.org/licenses/by/4.0/>



A Reproducible Codex Skill Workflow for Serbian Word Sense Disambiguation with Serbian WordNet

Saša Petalinkar¹[\[0009-0007-9664-3594\]](#), Nikola Vulović²[\[0009-0008-4260-348X\]](#)

¹University of Belgrade, Faculty of Mining and Geology, Serbia,
sasa.petalinkar@rgf.bg.ac.rs

²ENVI Software Solution, Belgrade, Serbia, nikolavulovic@hotmail.com

Abstract

Word sense disambiguation (WSD) for Serbian is difficult because rich inflection, productive derivation, multiword expression (MWE) targets, named entities, and limited sense-annotated data all increase the gap between surface forms and lexical-semantic interpretation. This extended abstract presents a reproducible Codex skill workflow for Serbian WSD with Serbian WordNet (SrpWN). The workflow combines Serbian part-of-speech (POS) and lemma preprocessing, SrpWN candidate retrieval, large language model (LLM)-based candidate comparison, structured JavaScript Object Notation (JSON) decision recording, and cautious extension suggestions. Compared with ordinary one-off prompting, the skill makes the annotation protocol explicit: exact phrase lookup is attempted for MWEs, lemma-based lookup is used for single-word targets, and every decision records the selected synset, rejected candidates, confidence, evidence, and possible inventory extension. A pilot evaluation was run over 250 sampled targets from WebAnno TSV data containing 2,024 texts, 31,059 tokens, 16,033 WSD annotations, 2,927 MWE-marked tokens, and 6,279 unique KB identifiers. The LLM-only pilot produced 191 linked items, 1 proposed extension aligned through Princeton WordNet (PWN 3.0) evidence, 51 proposed new extensions, and 7 skipped items; expert validation is outside the scope of the reported pilot evaluation.

Keywords: SrpWN, WSD, Codex skills, Serbian NLP, language resources, LLMs

1. Introduction

WSD is a central task in semantic natural language processing: given a word or phrase in context, a system must select the intended meaning from a sense inventory. For Serbian, this task is especially demanding. Serbian is morphologically rich, with extensive inflectional variation in nouns, adjectives, verbs, and pronouns. Lemma recovery is therefore non-trivial, and ambiguous surface forms may correspond to different lemmas, parts of speech, or syntactic roles. In addition, many lexical meanings are expressed by MWEs, named entities, or domain-specific terms, whose semantic behaviour is not reducible to the independent senses of their component words.

For lower-resourced languages, WSD is further constrained by the lack of large manually validated corpora and by the uneven coverage of sense inventories. Recent large language models (LLMs) offer useful semantic reasoning capabilities, but their use in WSD can remain difficult to reproduce if the annotation process is expressed only as a prompt. A prompt may contain useful instructions, but unless a prompt-only run is paired with explicit logging, it does not by itself preserve the full workflow: how candidate senses are retrieved, which linguistic evidence is trusted, which candidates were rejected, and how missing senses should be recorded. For example, if only a final synset identifier is saved, a later reviewer cannot tell whether a disagreement came from an absent candidate, a wrong lemma/POS cue, or a contextual choice among valid alternatives. The resulting annotations may be plausible, but they can be harder to audit or reuse.

This extended abstract presents a reproducible Codex skill workflow for Serbian WSD with SrpWN. The contribution is methodological: instead of treating LLMs as black-box annotators, the workflow packages preprocessing, lexical lookup, candidate comparison, output validation, and decision recording into an explicit skill. The skill makes the system behaviour inspectable and easier to repeat across batches of Serbian examples. It is designed for fast and reproducible LLM-only pilot annotation: in this paper, LLM-only means that no human annotator or expert adjudicator corrected individual target labels during the pilot run.

2. Codex Skill Workflow

Following the public Codex Agent Skills documentation, a Codex skill is a reusable directory-based package centred on a `SKILL.md` specification, with optional scripts, references, assets, and metadata that Codex can load when a task matches the skill (OpenAI, 2026). In this work, the Serbian WSD skill is therefore not just a prompt: it bundles the task specification, preprocessing and evaluation scripts, SrpWN lookup procedure, JSON/audit schema, and worked examples. Figure 1 shows the skill as a small versionable folder: the instructions, scripts, schema, lookup helpers, examples, and tests are stored together, so the annotation procedure can be rerun and inspected.

```

serbian-wsd-srpwn/
  SKILL.md                # task protocol
  references/
    srpwn_lookup_protocol.md  # lookup order
    annotation_examples.json  # worked examples
  scripts/
    prepare_judig_pilot_sample.py  # sample builder
    run_judig_pilot_evaluation.py  # pilot runner
  src/serbian_wsd_srpwn/
    schema.py                # JSON validation
    audit.py                 # wsd-audit-v1 records
    srpwn.py                 # SrpWN lookup
    english_wordnet.py       # PWN bridge
  tests/
    test_schema.py
    test_pilot_evaluation.py

```

Figure 1: Minimal structure of the Serbian WSD Codex skill package.

The skill is the reproducibility layer around LLM-assisted WSD: it specifies what evidence to gather, how to query the sense inventory, how to compare candidates, and how to store each decision. As Figure 2 shows, the workflow first preprocesses the Serbian text, then retrieves SrpWN candidate synsets, compares candidates against sentence context, records a JSON annotation and audit record, and marks unresolved cases as extension proposals.

In the Serbian WSD skill, preprocessing begins with sentence segmentation, tokenisation, and candidate identification. Serbian POS and lemma evidence is produced with the Hugging Face spaCy pipeline Tanor/sr_pln_tesla_j355, installed as the Python package `sr_pln_tesla_j355` version

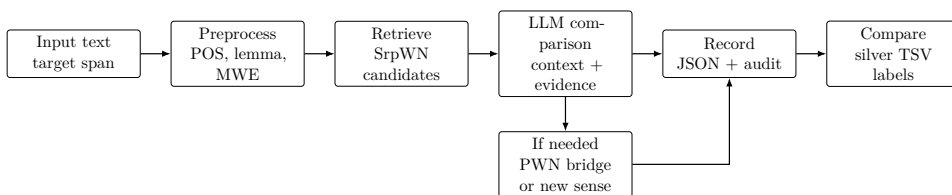


Figure 2: Workflow executed by the skill from Serbian input text to audit-oriented pilot comparison.

1.0.0 and loaded in code with `sr_pln_tesla_j355.load()`, with `spacy.load("sr_pln_tesla_j355")` as fallback (Ikonić Nešić et al., 2026). The model card reports spaCy compatibility $\geq 3.7.2, < 3.8.0$ and the pipeline components `transformer`, `tagger`, `trainable_lemmatizer`, and `ner`; Ikonić Nešić et al. (2024) describe and evaluate the related Serbian BERT/spaCy training context. This output is treated as evidence rather than gold annotation: automatic POS and lemma predictions are useful for candidate generation, but final WSD decisions must remain sensitive to context and to SrpWN evidence.

The skill separates single-word and multiword handling. For a single lexical target, the system considers the surface form, lemma, part of speech, and sentence context. For potential MWEs, surface n-grams and annotated phrase spans are checked first, because phrase meanings may correspond to a single SrpWN synset that would be missed by word-by-word lookup. Candidate synsets are retrieved from SrpWN and represented with the evidence needed for comparison: synset identifier, part of speech, Serbian literals, definition, domain information, usage examples when available, and relevant relations.

The LLM decision step compares candidates with local context and selects a synset only when definition and lexical evidence match. The pilot targets lexical content categories: nouns and proper nouns, verbs, adjectives, adverbs, and MWEs represented by these categories. The status `skipped` is used outside this scope, for example the preposition *u* in *u Beogradu*, punctuation, particles, or non-lexical tokens selected inside an MWE annotation. Remaining unresolved lexical items are marked as proposed extensions. The extension status is split into two cases. The status `proposed_extension_aligned` is used when no SrpWN candidate fits but a PWN synset can provide a

plausible PWN 3.0 anchor through a `suggested_eng30_id`. The status `proposed_extension_new` is used when neither SrpWN nor PWN provides a fitting sense. In both cases, PWN evidence is treated as review material, not as a replacement for Serbian lexicographic judgement.

The output is constrained as JSON. Figure 3 gives the complete linked annotation item used for the *merna jedinica* example below, including schema version, preprocessing evidence, lookup strategy, candidates, selected synset, confidence, evidence, and TSV comparison. For batch work, the skill saves a `wsd-audit-v1` record preserving input text, preprocessing, phrase candidates, SrpWN/PWN lookups, LLM corrections, final annotations, and extension suggestions. The audit layer supports review, disagreement analysis, and expert validation without reconstructing decisions from free-form dialogue.

3. Serbian WordNet as the Sense Inventory

Serbian WordNet (SrpWN) is used as the central sense inventory. In this workflow, SrpWN is not treated as a flat list of labels, but as a structured lexical-semantic resource whose synsets, literals, definitions, usage examples, domains, and relations provide evidence for disambiguation. This is important for Serbian, where inflection, derivation, and MWEs often obscure the relation between a surface form and its intended sense. Earlier work connected SrpWN development with Princeton WordNet, EuroWordNet, and BalkaNet-style multilingual wordnet construction, while later work explored corpus-based validation, cooperative enrichment, and resource-based augmentation (Krstev et al., 2004, 2008; Stanković et al., 2018).

The public SrpWN web interface at <https://wn.jerteh.rs/> allows searching by Serbian or English forms and by fields such as literals, definitions, usage examples, and domains. It also supports different lookup modes, including exact phrase search, begins-with search, and contains search (JeRTeh, 2026). The present pilot operationalises the visible web search behaviour as a lookup protocol rather than relying on a separately documented API: exact phrase lookup is attempted first for MWEs, followed by lemma-based lookup for single-word targets, while broader lookup is reserved for uncertain or failed exact searches. For reproducibility, the audit record stores the returned candidate evidence used for each decision, so later re-

```
{
  "schema_version": "wsd-audit-v1",
  "item_id": "judig-demo-merna-jedinica",
  "text": "Prema SI, merna jedinica sile je njutn (oznaka N).",
  "target": {
    "surface": "merna jedinica",
    "lemma": "merna jedinica",
    "pos": "n",
    "span": [9, 23],
    "mwe": true
  },
  "preprocessing": {
    "model": "Tanor/sr_pln_tesla_j355",
    "evidence": ["POS=NOUN", "lemma=merna jedinica"]
  },
  "lookup": {
    "inventory": "SrpWN",
    "strategy": ["exact_phrase", "lemma_literal"],
    "query": "merna jedinica"
  },
  "candidates": [
    {
      "synset_id": "ENG30-13583724-n",
      "literals": ["jedinica", "merna jedinica", "jedinica mere"],
      "domain": "metrology",
      "definition": "quantity accepted as a standard of measurement"
    }
  ],
  "decision": {
    "status": "linked",
    "selected_synset_id": "ENG30-13583724-n",
    "confidence": "high",
    "evidence": [
      "sentence refers to an SI unit",
      "exact MWE phrase is present in SrpWN",
      "domain and definition match measurement context"
    ]
  },
  "extension": null,
  "agreement_with_tsv": "yes"
}
```

Figure 3: Full JSON annotation item for the linked MWE example.

view does not depend on re-querying or re-scraping the live page. SrpWN remains the primary sense inventory; PWN 3.0, accessed locally through the Natural Language Toolkit (NLTK), is used only as supporting evidence for possible SrpWN expansion according to the extension procedure defined in the workflow section.

Figure 4 illustrates this candidate-generation step for the query *merna jedinica* in the SrpWN contains-search result view. The interface reports four returned Serbian and English synsets and exposes the first bilingual candidate, ENG30-13583724-n, with Serbian literals such as “jedinica”, “merna jedinica”, and “jedinica mere”, an English unit-of-measurement counterpart, and the metrology domain.

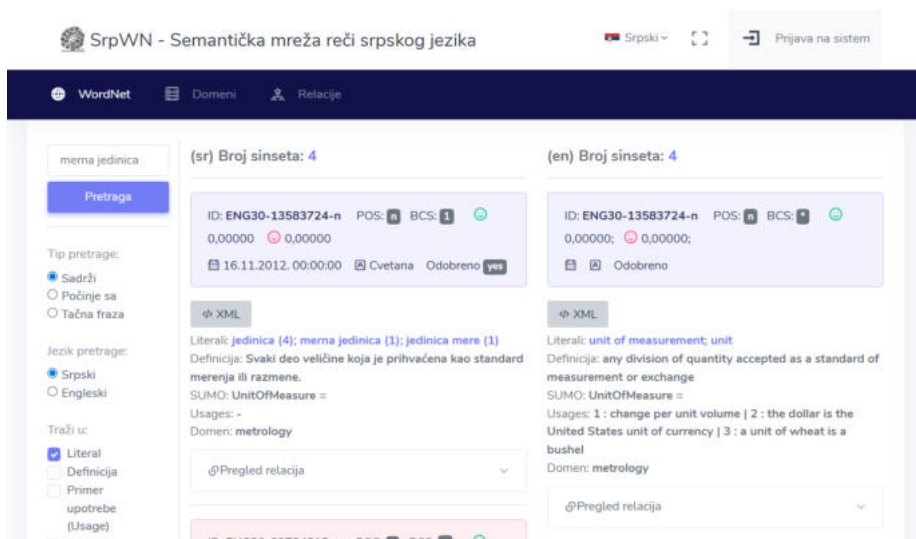


Figure 4: SrpWN contains-search page for the query *merna jedinica*, showing the full site frame and the first bilingual candidate synset returned during candidate generation. Source: <https://wn.jerteh.rs/>.

This resource-centric (sense-inventory-first) approach contrasts with context-only prompting and follows SR-ELEXIS annotation principles: MWEs and named entities are treated as semantic targets, labels are grounded in sense repositories, and lexicon-corpus links are prepared for Linguistic Linked Open Data representation (Krstev et al., 2024). The skill therefore does not replace SrpWN or expert lexicographic work; it preserves evidence for later review.

4. Annotation Examples

The annotation layer is intentionally more detailed than a final label column. It preserves the target span, the preprocessing evidence, the SrpWN search path, the selected or rejected candidate senses, and the reason for the final status. This makes it possible to inspect whether a disagreement comes from a real semantic alternative, from incomplete inventory coverage, from a lemma or POS problem, or from an MWE boundary decision. The first example below is also the case encoded in Figure 3.

One linked example comes from the sentence: “Prema SI, merna jedinica sile je njutn (oznaka N).” The target surface form is *merna* in the phrase *merna jedinica*. Exact phrase lookup in SrpWN returns ENG30-13583724-n, with Serbian literals including “jedinica” (“unit”), “merna jedinica” (“unit of measurement”), and “jedinica mere” (“unit of measure”); the definition describes a quantity accepted as a standard of measurement or exchange, and the domain is metrology. Because the sentence refers to the SI unit of force, this synset is selected and the item is recorded with status `linked`, selected synset ENG30-13583724-n, and evidence that the phrase sense matches the measurement context. The example also illustrates why phrase lookup is attempted before word-level lookup: analysing *merna* independently would lose the lexicalised unit-of-measure sense expressed by the full expression.

A contrasting proposed-extension example comes from the sentence: “Nakon te bitke borbe su se ponovo prenele na francusko tlo, dok je Venecija sklopila separadni mir.” The target surface form is *separadni*, with the lemma *separatan* and adjective POS evidence. In the pilot run, no fitting SrpWN candidate was available for the contextual meaning in *separadni mir*. Following the workflow section’s extension procedure, the item is recorded as `proposed_extension_aligned` because PWN returns *separate.a.01*, encoded as ENG30-02109678-a, for the English gloss “separate”. This is still only expansion evidence: the final Serbian literal, definition, and relation placement require later lexicographic review.

These examples show how the skill separates three tasks that are often blurred in prompt-only annotation: finding candidate senses, choosing a contextual sense, and identifying inventory gaps. They also show why every batch decision is saved as an audit record rather than only as a final label: later reviewers can see the candi-

date generation path and the justification for both linked and missing-sense decisions.

5. Data and LLM-Only Pilot Evaluation

The pilot uses existing ELEXIS-SR-WSD WebAnno TSV files prepared for Serbian WSD work and described in the submitted LLM-based Serbian WSD framework by [Petalinkar et al. \(2026\)](#), which serves here as the source description for the TSV data. These files contain Serbian WSD annotations produced with ChatGPT/LLM assistance and aligned with SrpWN-style identifiers; they are therefore treated here as silver, LLM-assisted comparison labels, not as expert-validated gold labels. The available dataset consists of five files covering 2,024 texts. A preliminary profiling pass over the TSV files found 31,059 tokens, 16,033 WSD annotations, 2,927 MWE-marked tokens, and 6,279 unique KB identifiers. The data include POS, lemma, MWE, and WSD layers, as well as fields for explanations, selected KB identifiers, number of candidate senses, origin, and possible sense identifiers. The profiling and all local data-processing steps are run through `uv run python` to keep execution reproducible inside the project environment.

The evaluation is explicitly LLM-only: no expert validation or human-in-the-loop correction/adjudication was performed during target labelling; human work was limited to workflow design, scripts, and aggregate/output inspection. The pilot compares the Codex skill output with silver TSV labels and simple baselines, such as first-candidate selection when available. The 250-target sample was stratified across MWEs, high-ambiguity targets, nouns and proper nouns, verbs, adjectives, adverbs, and extension-candidate targets, exposing the workflow to major sources of difficulty rather than maximising raw agreement.

The pilot measures are audit-oriented, and Table 1 summarises the main outputs. All 250 targets have valid audit JSON records. The status counts show that the workflow links most sampled targets while also exposing inventory gaps or retrieval limits: 191 items were linked, 52 were extension proposals, and 7 were skipped as outside the target scope. TSV comparison is identifier-based: yes means that the selected ENG30 synset matches a TSV possible/KB identifier, or that a TSV LLM/non-ENG30 label is also treated by the workflow as an inventory-gap/extension case; no means no such match; n/a is

reserved for skipped items. Different ENG30 IDs, related but non-identical IDs, and MWE span changes are not collapsed; they remain review cases.

Measure	Result	Interpretation
Audit JSON validity	250/250	All sampled targets produced parseable <code>wsd-audit-v1</code> records.
Linked decisions	191	Most targets could be attached to an SrpWN/PWN-style identifier.
Proposed extensions	52	Missing or non-fitting SrpWN coverage remained visible for later lexicographic review.
Skipped items	7	Items outside the target lexical categories were excluded from agreement scoring.
Agreement with TSV label	211 yes, 32 no, 7 n/a	Comparison is against silver LLM-assisted TSV labels, not expert gold labels.
Extension-candidate bridge	1/20	PWN provided one anchor among the extension-focused subset, so the bridge is useful but conservative.

Table 1: LLM-only pilot results over 250 sampled WSD targets.

The extension-candidate subset had 20 items and one PWN bridge anchor (5%), as shown in Table 1. For each non-agreement, the review prompt asks the LLM to compare the TSV label, selected or available SrpWN candidates, lemma/POS cue, MWE span, and any PWN anchor, then choose one category: alternative plausible sense, missing SrpWN sense, lemma or POS problem, MWE boundary issue, PWN alignment mismatch, or other lookup mismatch, with a short evidence note.

6. Discussion and Conclusion

The proposed workflow contributes to Serbian WSD in two ways. First, it offers a practical method for applying LLMs to a lower-resourced, morphologically rich language while keeping the annotation process explicit. LLMs are not asked simply to “choose a meaning”; they are required to follow a documented process, use SrpWN as a structured sense inventory, record the evidence for each

decision, and identify cases where the inventory may need extension. Second, the workflow helps bridge automatic annotation and lexicographic resource development. Proposed extensions, MWE boundary issues, lemma/POS corrections, and possible PWN alignment anchors are not lost as incidental comments, but stored as structured data that can later be reviewed.

The main limitation is deliberate: the extended abstract reports the method and an LLM-only pilot run, not expert-validated accuracy. The reported numbers should therefore be read as reproducibility and audit evidence, not as final WSD accuracy. Future evaluation should add expert validation, adjudication of disagreement categories, and a stronger comparison with embedding-based or first-sense baselines. Even with this limitation, the workflow demonstrates how Codex skills can support transparent and reusable semantic annotation for Serbian and similar lower-resourced languages.

Language Resources

Ikonić Nešić, Milica and Petalinkar, Saša and Škorić, Mihailo and Stanković, Ranka. 2026. *Tanor/sr_pln_tesla_j355*. Hugging Face. Model card available at https://huggingface.co/Tanor/sr_pln_tesla_j355; for Serbian POS tagging, lemmatisation, NER; Python package *sr_pln_tesla_j355* version 1.0.0.

JeRTeh. 2026. *Serbian WordNet: Semantic Network of Serbian Words*. JeRTeh – Society for Language Resources and Technologies. PID <https://wn.jerteh.rs/>.

References

Milica Ikonić Nešić, Saša Petalinkar, Mihailo Škorić, and Ranka Stanković. 2024. [BERT Downstream Task Analysis: Named Entity Recognition in Serbian](#). In *Disruptive Information Technologies for a Smart Society*, volume 860 of *Lecture Notes in Networks and Systems*, pages 333–347. Springer.

Cvetana Krstev, Gordana Pavlović-Lažetić, Duško Vitas, and Ivan Obradović. 2004. [Using Textual and Lexical Resources in Developing Serbian Wordnet](#). *Romanian Journal of Information Science and Technology*, 7(1–2):147–161.

- Cvetana Krstev, Ranka Stanković, Aleksandra M. Marković, and Teodora Sofija Mihajlov. 2024. [Towards the Semantic Annotation of SR-ELEXIS Corpus: Insights into Multiword Expressions and Named Entities](#). In *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024*, pages 106–114, Torino, Italia. ELRA and ICCL.
- Cvetana Krstev, Bojana Dorđević, Sanja Antonić, Nevena Ivković-Berček, Zorica Zorica, Vesna Crnogorac, and Ljiljana Macura. 2008. [Cooperative Work in Further Development of Serbian Wordnet](#). *Infotheca – Journal of Informatics and Librarianship*, 9(1/2):59a–78a.
- OpenAI. 2026. Agent Skills – Codex. <https://developers.openai.com/codex/skills>. Accessed 28 May 2026.
- Saša Petalinkar, Ranka Stanković, Milica Ikonić Nešić, and Cvetana Krstev. 2026. A Semi-Automated LLM-Based Framework for Word Sense Disambiguation in Serbian. Submitted manuscript.
- Ranka Stanković, Miljana Mladenović, Ivan Obradović, Marko Vitas, and Cvetana Krstev. 2018. [Resource-based WordNet Augmentation and Enrichment](#). In *Proceedings of the Third International Conference on Computational Linguistics in Bulgaria (CLIB 2018)*, pages 104–114, Sofia, Bulgaria. Department of Computational Linguistics, Institute for Bulgarian Language, Bulgarian Academy of Sciences.

Reproducibilni Codex Skill tok rada za razrešavanje višeznačnosti reči u srpskom jeziku pomoću Srpskog WordNeta

Saša Petalinkar, Nikola Vulović

Sažetak

Rad predstavlja tok rada zasnovan na Codex veštini za WSD u srpskom jeziku pomoću SrpWN-a. Cilj nije da se prikaže konačna ekspertno validirana evaluacija, već da se opiše ponovljiv i proverljiv metod za LLM-potpomognutu semantičku anotaciju. Tok rada kombinuje predobradu teksta, POS i lema informacije kao pomoćne dokaze, pretragu SrpWN kandidata, poređenje kandidata u kontekstu i beleženje odluka u strukturisanom JSON formatu. Kada u SrpWN-u ne postoji odgovarajući kandidat, PWN 3.0 se koristi samo kao pomoćni dokaz za moguće proširenje inventara, a ne kao zamena za SrpWN ili ekspertsku leksikografsku procenu. Pilot evaluacija obuhvatila je 250 uzorkovanih ciljeva i proizvela 191 povezanu oznaku, 1 predlog proširenja poravnat pomoću PWN dokaza, 51 predlog novog proširenja i 7 preskočenih stavki. Ekspertska validacija, adjudikacija neslaganja i potpunija statistička analiza navedene su kao budući nastavak evaluacije, a ne kao deo ovde prikazanog LLM-only pilota.

Ključne reči: Srpski WordNet, razrešavanje višeznačnosti reči, Codex veštine, srpski jezik, jezički resursi, veliki jezički modeli

Works are licensed under the terms and conditions of the Creative Commons Attribution 4.0 International License.

<https://creativecommons.org/licenses/by/4.0/>



Графичната организация на текста като аналитичен обект при изграждането на диахронен корпус

Таня Нейчева¹[0000-0003-0117-8524]

¹ Пловдивски университет „Паисий Хилендарски“,

Филологически факултет (neycheva@uni-plovdiv.bg)

Резюме

Докладът обсъжда от гледна точка на потребителя на езикови корпуси някои проблеми, свързани с дигиталното представяне и последващата обработка на славянски ръкописни текстове от преднационалната и ранната национална епоха. По-специално, това са графичните особености на ръкописната традиция – нестандартизирана сегментация на текста, надредни знаци, съкращения под титла, преноси и други елементи, които често се губят при нормализацията и корпусната обработка.

Задава се въпросът дали е практически възможно създаването на многослоен модел на дигитално представяне, който да съчетава текстологичен с традиционния лингвистичен слой (или слоеве). Основната цел на подобен текстологичен слой би била съхраняването на структурирана и машинно обработваема информация за графичната организация на ръкописа наред с данните, необходими за езиков анализ.

Информацията за материалната реализация на текстовете от преднационалната и ранната национална епоха, ако бъде достъпна за автоматична обработка, търсене и последваща анотация, би била изключително ценен източник за всички изследователи на историята на книжовната норма, а също и за целите на текстологията.

Ключови думи: диахронен корпус, многослоен корпус, ръкописи, текстология

1. Увод

Дигиталното представяне на славянските ръкописи (български, църковнославянски и др.) от преднационалната и ранната национална епоха поставя редица трудности. Проблемно е както

представянето и интерпретирането на изнесените над реда букви, така и сегментирането на текста в съответствие със съвременните представи за границите на лексемата.

Настоящият коментар е опит на един *потребител* на дигиталните корпуси да предложи структурен модел на многослоен дигитален корпус от ръкописни паметници, който да съчетава максимално близко до оригинала визуално представяне с нормализация и последваща анотация. Целта е да се създаде тексто-логичен слой, който да съхранява в машиннообработваем вид информация за материалната и графичната организация на писмения текст наред с неговото езиково съдържание.

2. Графичната организация на ръкописния текст и нейното дигитално представяне

Под „графична организация“ тук разбираме съвкупността от средства, чрез които писаният текст структурира и предава езиковата информация – не само на равнище буквен запис, но и чрез диакритичните и другите надредни знаци, които са извън линейната последователност, чрез слятото и разделното изписване на думите, чрез преносите на нов ред и нова страница и други текстови маркери. Тези елементи от материалната страна на ръкописа образуват система, която пази немаловажна информация за писмените практики на съответната книжовна традиция.

При дигиталното представяне всяко ниво на допълнителна нормализация или структуриране на текста в съответствие със съвременните разбирания неизбежно променя или премахва част от оригиналната информация. По-долу са разгледани някои от основните проблемни зони.

2.1. Сегментация на текста

В периода от най-ранните славянски ръкописни паметници и до края на XV в. обичайна практика е текстът да се изписва слято, без разстояние между думите. Границите между думите не са обозначени в оригинала и се интерпретират от четящия. Едва през XV–XVI в. с появата на първите печатни книги¹ на кири-

¹ През 1491 г. Швайполт Фиол отпечатва в Краков първите кирилски книги – „Октоих“ и „Часослов“. В тях разстояния между думите липсват, т.е. пре-създадена е ръкописната традиция. Почти по същото време обаче в Цетине вече е отпечатана книга, която съдържа интервали – „Псалтир“ (1495).

лица под влияние на западноевропейската типографска практика постепенно се разпространява и усвоява употребата на интервали в печата и в ръкописите. В този смисъл графичната организация на текста преди споменатата промяна носи смисъл, който не съвпада със съвременните стандарти.

В по-късната традиция (XVI – XVIII в.) постепенно се утвърждава разделянето на текста на фонетични думи, състоящи се най-често от една лексема със собствено ударение и една или няколко клитики. Нещо повече, употребата на диакритичните знаци, характерни за писмеността през този период, се подчинява на мястото и ролята им именно в така обособената фонетична дума.

Още първите научни издания на ръкописи през XIX в. обаче въвеждат практиката текстът да бъде сегментиран на отделно изписани думи в съответствие със съвременните разбирания за границите на лексемата и в нарушение на оригиналната сегментация. Тази традиция най-често се запазва и днес при дигитализирането на славянското книжовно наследство (изключение прави *Ground Thruth* слойт в корпуса от славянски текстове на [Staatsbibliothek zu Berlin \(2025\)](#)), като така значително се улеснява не само зрителното възприемане от съвременния читател, но и автоматичната обработка на текста.

Недостатък на сегментирането е това, че то неизбежно се извършва от този, който обработва текста – било то чрез съпоставка с предварително зададен речник, било чрез индивидуална интерпретация. И в двата случая в текста се въвежда външно решение, което в отделни случаи може и да не съвпада с авторовото (или на преписвача) възприемане на границите между думите. В резултат съвременната сегментация е в състояние да промени не само графичния облик на текста, но и начина, по който той се интерпретира.

2.2. Съкращения под титла

Съкращенията под титла, обикновена или буквена, изпълняват различни функции в текста: обозначават понятия със сакрален характер, използват се за икономия на място или сигнализируют, обикновено в комбинация с пунктуационни знаци, числовата стойност на кирилските букви. При дигиталното представяне на паметниците се наблюдават различни подходи към предаването на титлите.

Понякога текстовете се дигитализират не по оригиналния ръкопис, а по по-късни археографски или литературоведски печатни издания поради недостъпност на оригинала или други причини. В някои случаи графичното оформление е до голяма степен запазено, в други обаче става въпрос за адаптиран текст, преминал през различни етапи на нормализация, която включва освен всичко друго и разкриване на съкращенията под титла. Дигитализацията на такива източници значително улеснява последващите стъпки на обработката и езиковия анализ, но води до заличаване на част от оригиналната графична информация. Немалък брой подобни текстове са включени например в староруския подкорпус на [НКРЯ \(2026\)](#).

Добрата практика на археографските издания изисква адекватно предаване на графичната организация на оригинала, включително запазване на съкращенията под титла, което е текстологически и палеографски коректно. Липсата на нормализация обаче в този случай затруднява търсенето в дигитализирания текст и автоматичната му обработка. В някои достъпни онлайн дигитални колекции това води до ограничена функционалност при търсене ([Korpus českocírkevněslovanských textů, 2025](#); [Cyrillo-methodiana, 2021](#)).

Тези затруднения до голяма степен се преодоляват чрез многослойно представяне на текста, при което се съхраняват както дипломатичен слой, максимално близък до оригинала, така и нормализиран слой, предназначен за търсене и автоматична обработка. Дипломатичният слой обаче е насочен предимно към визуалното възпроизвеждане на текста и не позволява самостоятелно изследване на графичната организация на писмения паметник.

2.3. Изнесени над реда букви

Изнесените над реда букви изпълняват различни функции в текста: използват се при съкращаване на думата, за поправка на грешки, за избягване на пренос на нов ред и др.

При научното представяне и в частност при дигитализирането на ръкописите съществуват няколко основни начина за тяхното предаване. Един от тях е буквата да бъде „свалена“ на реда без допълнителни означения. Това е обичайна практика например в историческите подкорпуси на [НКРЯ \(2026\)](#), при което се нарушава поне една от правописните особености на оригинала,

тъй като ако над реда е изнесено окончание, то редовно е свързано и с пропуск на краесловния ер, който е задължителен, ако окончанието не е изнесено. Преместването на изнесените букви в линейния ред води до заличаване на тази графико-правописна опозиция.

Друг разпространен начин на предаване, наследен от печатните археографски издания, е при „свалянето“ буквата да бъде обозначена чрез затваряне в скоби. Като илюстрация могат да послужат някои текстове, дигитално представени в [Scripta Bulgarica \(2014\)](#) и [Korpus českocírkevněslovanských textů \(2025\)](#). Този подход променя графичната структура на текста чрез въвеждане на допълнителни символи (скоби) и може да затрудни последващата обработка.

И двата описани подхода до голяма степен следват логиката на линейното текстово оформяне. По-близо до оригиналната графика изглежда решението, при което се запазва изписването над реда при визуализацията (каквато е практиката в текстовия корпус на [Cyrillomethodiana \(2021\)](#)). Там е решен успешно и въпросът с търсенето по словоформа, без да се налага въвеждането на допълнителни символи, т.е. основната буква и изнесената буква се тълкуват като еквивалентни. Недостатък на подобно представяне е, че данните за това кога и как се употребяват изнесени букви не могат да бъдат извлечени в систематизиран вид за последващ анализ.

2.4. Диакритични знаци

Макар че, за разлика от гръцката писменост, ударенията и придиханията в славянската традиция нямат различни фонетични стойности, тяхната употреба не е хаотична – тези знаци следват правописни правила, които може да варират в зависимост от книжовната школа, времето на създаване на текста или индивидуалния навик на пишещия. Именно затова тяхното системно представяне в дигиталния текст и възможността за последващата им анотация биха могли да бъдат полезни както за езикови, така и за текстологични изследвания.

Диакритичните знаци обикновено се съхраняват при дигиталното представяне на ръкописите (както и при книжните научни издания), но в онлайн ресурсите най-често остават част предимно от визуализацията на текста, а не на носената от не-

го структурирана и машинно обработваема информация. Те в повечето случаи не са достъпни за търсене и анализ като самостоятелен обект на изследване. Това от своя страна затруднява проследяването на принципите и закономерностите, свързани с употребата им.

2.5. Пренос на нов ред и нова страница

Предвид факта, че вече са налични дигитални изображения на голяма част от ръкописите, при това с много добро качество, точното следване на оформлението на страницата при представянето им в текстов формат до голяма степен е изместено на заден план. Преносът на една дума на нов ред обаче не е само механично разпределение на текста върху страницата, а по необходимост следва правописни правила, които могат да варират според книжовната традиция или индивидуалната писателска практика. В този смисъл елиминирането на преносите може да доведе до загуба на информация, която, макар и минимална, не е без значение за графичния и правописния анализ.

В печатните издания и при дигиталното представяне могат да се обособят два основни подхода към проблема за пренесените думи. Първият, характерен за редица корпусни ресурси ([НКРЯ 2026](#); [Korpus českocírkevněšlovanských textů 2025](#); [Cyrillomethodiana 2021](#) и др.), предполага отпадане на преносите, а понякога и на странирането с цел улесняване на последващото търсене или анотация. В резултат част от носената от материалната страна на текста информация не се съхранява в аналитично достъпен вид. Вторият подход включва запазването на преносите и оригиналното деление на текста на редове и страници с цел по-адекватно визуално и текстологично представяне ([Scripta Bulgarica 2014](#); [Staatsbibliothek zu Berlin 2025](#); [Korpus českocírkevněšlovanských textů 2025](#) и др.), но не е достъпен като аналитичен слой, от който може да бъде извлечена обобщена информация.

2.6. Графичната организация на текста като аналитичен слой

Изброените по-горе бележки се основават не само на наблюдения върху съществуващите онлайн диахронни ресурси, но и на личен опит при изграждането на малък корпус, включващ руски ръкописни и печатни паметници от XVIII в., в който тек-

стът беше частично нормализиран за нуждите на морфологичната анотация чрез сегментиране според съвременните представи, премахване на диакритичните знаци и др. В хода на тази работа се появиха редица трудности, свързани със загубата на графична и правописна информация и с избора между различни нива на представяне на текста. От гледна точка на потребител на дигиталните корпуси, който лично се е сблъскал с проблемите, авторът на този доклад предлага идея за разработване на допълнителен аналитичен слой в корпусното представяне на текстове от преднационалната епоха, който да е посветен именно на графичната информация.

3. Идеи за моделиране на графичната информация като отделен слой в диахронен корпус

В основата на един диахронен корпус лежи ръкописът или старопечатната книга като материален обект и източник на дигиталните изображения, които, от своя страна, са източник на текстовите данни. За да бъде запазено максимално количество от разнообразната информация, предоставяна от оригинала, текстът следва да бъде представен в няколко успоредни слоя, между които да се установи еднозначно съответствие, позволяващо паралелното им представяне и автоматичното преминаване от един слой в друг.

Един евентуален текстологичен слой би имал за цел не само и не толкова за адекватната визуализация (за която може да бъде използвано с много по-голям успех изображението), а в машинно обработваем вид. Неговата задача не е да улесни четенето или езиковия анализ, а да съхрани възможно най-много от особеностите на графичната организация на текста: надредни знаци, съкращения, особености на сегментацията и други графични маркери. На фиг. 4 е показан пример за съответствието между изображението и текстологичния слой.

3.1. Оригинална сегментация на текста

Съвременната ръкописна норма разполага с един символ, обозначаващ границата между думите – празно разстояние, интервал, чиято ширина може да варира в зависимост от почерка на пишещия. В текстовете от XVI–XVIII в. обаче се употребяват най-малко два типа разстояния. Между отделните фоне-

тични комплекси обикновено се наблюдава ясно изразен интервал, понякога широк колкото две стандартни букви, а също така и значително по-малки разстояния, обикновено по-тесни от една буква, които визуално отделят някой елемент, осъзнаван от пишещия като клитика, без да нарушават единството на фонетичния комплекс. Разликата между двата типа интервали, макар и размерът им да не може да бъде съвсем точно дефиниран, изглежда последователна и носи информация за начина, по който преписвачът структурира текста.

За да бъде съхранено това разграничение, в текстологичния слой, би могло двата типа интервали да се отбелязват с различни знаци. Пълният интервал впоследствие да се използва като граница, отделяща една от друга фонетичните думи, всяка от които да бъде разглеждана като отделен токен, а непълният интервал да остава част от същия токен. По този начин би могла да се запази информацията за оригиналната сегментация на текста, без да се затруднява последващата обработка.²

3.2. Съкращения под титла и изнесени над реда букви

Разкриването на думите под титла в текстологичния слой се приема не изглежда целесъобразно, те трябва да се запазят във вида, в който са представени в оригинала.

Изнесените букви нямат специфична семантика, която да ги отличава от основния вариант на графемата, а информацията, която носят, е най-вече от текстологичен характер и би била от полза при изследване на индивидуалните правописни навици на книжовника. Тъй като тези букви са кодирани като самостоятелни Unicode символи, машинното им представяне не създава трудности само по себе си. Трудности възникват в случаите, когато при изнасянето в краесловие се пропуска краен елемент от окончанието (най-често, но не единствено краесловният ер), който се подразбира от контекста. Този тип явления не би трябвало да се решават на нивото на текстологичния слой. Неговата задача е да остане максимално близък до оригинала и да не въвежда редакторски интервенции под формата на „възстановяване“ на пропуснати елементи, нормализация или затваряне в скоби.

² Ако по аналогична методика бъде представен източник, в който текстът е изписан без разстояния между думите, той би трябвало да се сегментира на по-големи отрязъци, например от пунктуационен знак до пунктуационен знак.

3.3. Диакритични знаци

Ударенията и придиханията, характерни за графиката на кирилските ръкописи след XV век (църковнославянски, български, сръбски и др.), не просто сочат мястото на ударението, но могат и да сигнализират начало или край на думата, а в някои случаи разграничават омоними или различни граматични форми на една лексема (Плетнева and Кравецкий, 2001). В този смисъл те не трябва да бъдат пренебрегвани и въпреки възможните трудности е добре да бъдат коректно предадени в текстовия слой. Чест проблем обаче при ръкописните текстове и в областта на диактириците е липсата на единна норма. За адекватното им представяне в корпуса вероятно трябва да бъдат предвидени достатъчно голям брой знаци, които да отразяват максимално близко оригинала, без да го подлагат на тълкуване.

Особен случай представлява кратката (бреве) над и, която в периода XVI–XVIII в. заема междинно място между диакритичен знак и елемент от графемата ѝ и се употребява непоследователно не само в различите паметници, но и в рамките на един и същи паметник. Поради тази причина в текстологичния слой би било логично случаи да бъдат регистрирани като комбинация и + кратка, а едва на по-късен етап в следващите слоеве отделните случаи да бъдат подложени на анализ и интерпретирани по един или друг начин.

3.4. Преноси на нов ред и нова страница

Преносите на думи, независимо дали са оформени с тире или по друг начин, би трябвало да се отбележат в текстологичния слой с различни символи, които да не разкъсват фонетичната дума, т.е. да останат вътре в границата на токена. При определянето на това дали една фонетична дума е свършила преди края на реда или е пренесена на следващия най-сигурният белег е наличието на тире. В останалите случаи, за съжаление, трябва да се разчита на външно решение, основаващо се на съответствие с предварително зададен речник или други показатели (например употребата на някои диакритични знаци).

3.5. Корекция на допуснати грешки

И още една особеност, която не беше обсъждана по-горе. В ръкописната традиция в случай на корекции книжовникът зачертава неправилно изписани думи или части от тях. Тези графични следи не следва да се премахват от текстологичния слой, тъй като носят допълнителна информация извън самия езиков материал – например за уменията на преписвача да забелязва и поправя грешки, за умора или моментна разсеяност, както и за възможно погрешно разчитане на текста при преписване. Запазването им със специални знаци и маркирането им чрез съответни тагове в анотацията би обогатило възможностите за по-задълбочен текстологичен анализ.

4. Заключение

Така създаденият текстологичен слой може да бъде снабден с метаданни за източника и анотиран чрез набор от тагове, описващи основните графични явления в текста. Към всеки токен може да бъде добавена информация за наличието в него на непълните интервали между думите, преноси на нов ред, диакритични знаци – наличие и позиция в думата, както и наличие на всякакви други графични особености. Подобна анотация би дала възможност графичните особености да бъдат изследвани както самостоятелно, така и в съчетание помежду им.

Language Resources

- Cyrrillomethodiana. 2021. *Исторически корпус и речници на българския език (HistDict)*. Софийски университет „Св. Климент Охридски“.
- Korpus českocírkevněslovanských textů. 2025. *Korpus českocírkevněslovanských textů*. Univerzita Palackého v Olomouci.
- Scripta Bulgarica. 2014. *Scripta Bulgarica: дигитална библиотека за старобългарска книжнина*.
- Staatsbibliothek zu Berlin. 2025. *Kirchenslavica: Textkorpus der kirchenslawischen Drucke*. (Slav)CorpHub.
- НКРЯ. 2026. *Национальный корпус русского языка*. Российская академия наук.

References

- А. А. Плетнева and А. Г. Кравецкий. 2001. *Церковнославянский язык: Учебник для общеобразовательных учебных заведений, духовных училищ, гимназий, воскресных школ и самообразования*. Древо добра, Москва.

The graphic features of manuscript texts as an analytical object in the construction of a diachronic corpus: a bulgarian damaskin case study

Tanya Neycheva

Summary

The paper discusses, from the perspective of users of language corpora, some problems related to the digital representation and subsequent processing of Slavic manuscript texts from the pre-national and early national periods. In particular, these include the graphic features of the manuscript tradition: non-standardized text segmentation, superscript signs, contractions, word division across line breaks, and other elements that are often lost in the process of normalization and corpus processing.

The paper raises the question of whether it is practically feasible to create a multilayer model of digital representation that combines a textual-critical layer with the conventional linguistic layer (or layers). The main purpose of such a textual-critical layer would be to preserve structured, machine-processable information about the graphic organization of the manuscript alongside the data required for linguistic analysis.

If information about the material realization of texts from the pre-national and early national periods were made available for automated processing, searching, and subsequent annotation, it would constitute an exceptionally valuable resource for all researchers studying the history of the literary norm, as well as for the purposes of textual scholarship.

Keywords: diachronic corpus, multilayer corpus, manuscript studies, textual criticism

Works are licensed under the terms and conditions of the Creative Commons Attribution 4.0 International License.

<https://creativecommons.org/licenses/by/4.0/>



Korpus javnog diskursa u italijanskom i srpskom jeziku: struktura, metodologija i istraživački potencijal

Jovana Bazić¹[0009-0004-3220-5542]

¹ Univerzitet u Kragujevcu, Filološko-umetnički fakultet, Centar za naučnoistraživački rad, Srbija, jovana.bazic@filum.kg.ac.rs

Apstrakt

U okviru istraživanja za potrebe doktorske disertacije „Ogradi- vanje u javnom diskursu: kontrastivna analiza italijanskog i srpskog jezika” (Bazić, 2025) formiran je uporedni (komparativni) korpus koji obuh- vata pisani i govorni javni diskurs u italijanskom i srpskom jeziku. Cilj je bio obezbeđivanje reprezentativnog i autentičnog uzorka savremenog jezika kao polazišta za kontrastivna istraživanja. Iako postoje različiti korpusi italijan- skog kao drugog ili stranog jezika (LIPS, BADIP, VIP), korpus ITALSERB za srbofone učenike (Ceković, 2022), paralelni književni korpus *SerbItaCor* na platformi *NoSketchEngine* (Moderc et al., 2023; Moderc, 2025), kao i novinski korpusi poput italijanskih uvodnika (Koprivica Lelićanin, 2014) ili postojećih novinskih resursa za srpski jezik (*SrWaC*, *Classla V2*, *ParCo- Jour*), do sada nije realizovan uporedni korpus javnog diskursa koji obuh- vata oba predmetna jezika i oba modaliteta (pisani i govorni). Iz tog razloga formiran je *Corpus of Italian and Serbian Public Discourse (CISPD)* kao uporedni korpus tematski usaglašenih, izvornih tekstova u oba jezika.

Relevantnost i aktuelnost jezičke građe obezbeđene su odabirom medijskih tekstova i emisija nastalih u periodu 2023–2024. godine. Pisani deo korpusa prikupljen je kombinovanim postupkom (automatskim preuz- imanjem/scraping i ručnom selekcijom) sa vodećih portala: za italijan- ski su to *Corriere della Sera*, *La Repubblica*, *Il Sole 24 Ore* i *Il Mes- saggero*, a za srpski *Danas*, *Politika*, *Blic* i *NIN*. Kriterijumi odabira bili su reprezentativnost, autentičnost, heterogenost sadržaja i tematska us- aglašenost (Tognini-Bonelli, 2001). Pisani potkorpus obuhvata po 250.000 reči za svaki jezik i pokriva politiku, ekonomiju, kulturu, zdravlje, umetnost i društvena pitanja.

Govorni deo prikupljen je sa televizijskih kanala i podkasta u oba jezika (po 250.000 reči). Za italijanski su obuhvaćeni *Rai*, *Sky*, *Mediaset*, *Nove*, *La7*, te podkasti *Passa dal Basement* i *Il Bazar Atomico*. Za srp-

ski jezik odabran je veći broj televizijskih kuća (*RTS, NovaS, N1, Prva, Insaider, BrainzTV, Hram*, uz podkaste *Agelast* i *Lobotomija*) kako bi se nadomestila razlika u produkcijskom obimu i obezbedila tematska i žanrovska ujednačenost sa italijanskim uzorkom. Audio-materijal je transkribovan kombinacijom automatskih alata za prepoznavanje govora i obimne ručne verifikacije i korekcije, čime je obezbeđena visoka tačnost transkripta. Svi tekstovi sistematizovani su u Word i Excel formatima. Reprezentativnost je potkrepljena i podacima o gledanosti/posećenosti dobijenim od zvaničnih agencija (*Audiweb, Audipress, Auditel, Accertamenti diffusione stampa, Gemius, Medija centar* i *Kliping*).

Formirani CISPD korpus trenutno broji ukupno oko 500.000 reči (uz planirano proširenje na milion reči građom iz 2025. i 2026. godine). Njegova uporedna (*comparable*) struktura pruža stabilnu platformu za pragmatičke, semantičke, sociolingvističke i diskursne analize, kao i primenu u nastavi i oblasti jezičkih tehnologija. Korpus je predviđen za automatsku morfosintaktičku obradu (npr. CLASSLA paketom), a po okončanju radova biće javno dostupan istraživačkoj zajednici putem otvorenih digitalnih platformi (link ka probnom uzorku dostupan je na zahtev).

Ključne reči: CISPD (Corpus of Italian and Serbian Public Discourse), italijanski jezik, srpski jezik, uporedni korpus, javni diskurs, govorni diskurs, pisani diskurs

References

- Jovana Bazić. 2025. *Ograđivanje u javnom diskursu: kontrastivna analiza italijanskog i srpskog jezika*. Ph.D. thesis, Univerzitet u Beogradu, Filološki fakultet, Beograd. Neobjavljena doktorska disertacija [Hedging in Public Discourse: A Contrastive Analysis of Italian and Serbian. Unpublished doctoral dissertation].
- Nevena Ceković. 2022. *Italijanski razgovorni jezik*. Filološki fakultet, Beograd.
- Marija Koprivica Lelićanin. 2014. *Metadiskurs u savremenim italijanskim akademskim i novinskim člancima*. Ph.D. thesis, Univerzitet u Beogradu, Filološki fakultet. Neobjavljena doktorska disertacija.
- Saša Moderc. 2025. [Korpus srpskog književnog jezika SerbItaCor3](#). *Srpski jezik*, XXX:257–267.
- Saša Moderc, Ranka Stanković, Aleksandra Tomašević, and Mihailo Škorić. 2023. [An Italian-Serbian Sentence-Aligned Parallel Literary](#)

Corpus. *Review of the National Center for Digitization*, (43):78–91.
Elena Tognini-Bonelli. 2001. *Corpus Linguistics at Work*, volume 6
of *Studies in Corpus Linguistics*. John Benjamins Publishing Com-
pany, Amsterdam and Philadelphia.

Corpus of Public Discourse in Italian and Serbian: Structure, Methodology, and Research Potential

Jovana Bazić

Sažetak

Within the research conducted for the doctoral dissertation *"Hedging in Public Discourse: A Contrastive Analysis of Italian and Serbian"* (Bazić, 2025), a comparative corpus covering written and spoken public discourse in Italian and Serbian was created. The objective was to provide a representative and authentic sample of contemporary language as a starting point for contrastive studies. Although various corpora of Italian as a second or foreign language exist (LIPS, BADIP, VIP), as well as the ITALSERB corpus for Serbian learners (Ceković, 2022), the parallel literary corpus SerbItaCor on the NoSketchEngine platform (Moderc et al., 2023; Moderc, 2025), and news corpora such as Italian editorials (Koprivica Lelićanin, 2014) or existing newspaper resources for Serbian (SrWaC, Classla V2, ParCoJour), a comparative public discourse corpus encompassing both target languages and both modalities (written and spoken) has not been implemented until now. For this reason, *the Corpus of Italian and Serbian Public Discourse (CISPD)* was constructed as a comparative corpus of thematically aligned, original texts in both languages.

The relevance and currency of the linguistic data were ensured by selecting media texts and broadcasts produced between 2023 and 2024. The written section of the corpus was collected through a combined approach (web scraping and manual selection) from leading portals: for Italian, *Corriere della Sera*, *La Repubblica*, *Il Sole 24 Ore*, and *Il Messaggero*; and for Serbian, *Danas*, *Politika*, *Blic*, and *NIN*. The selection criteria included representativeness, authenticity, content heterogeneity, and thematic alignment (Tognini-Bonelli, 2001). The written subcorpus comprises 250,000 words per language, covering politics, economics, culture, health, art, and social issues.

The spoken section was gathered from television channels and podcasts in both languages (250,000 words each). The Italian dataset includes *Rai*, *Sky*, *Mediaset*, *Nove*, and *La7*, alongside the podcasts *Passa dal Basement* and *Il Bazar Atomico*. For Serbian, a larger number of TV outlets was selected (*RTS*, *NovaS*, *N1*, *Prva*, *Inšajder*, *BrainzTV*, *Hram*, along with the podcasts *Agelast* and *Lobotomija*) to compensate for differences in production volume and ensure thematic and genre consistency with the

Italian sample. Audio materials were transcribed using automated speech recognition tools combined with extensive manual verification and editing, ensuring high transcript accuracy. All texts are organized in Word and Excel formats. Representativeness is further supported by viewership and readership data provided by official agencies (*Audiweb*, *Audipress*, *Auditel*, *Accertamenti diffusione stampa*, *Gemius*, *Medija centar* and *Kliping*).

The established CISPD corpus currently totals approximately 500,000 words (with a planned expansion to one million words using data from 2025 and 2026). Its comparable structure offers a robust platform for pragmatic, semantic, sociolinguistic, and discourse analysis, as well as applications in education and language technologies. The corpus is prepared for automated morphosyntactic annotation (e.g., using the CLASSLA pipeline) and will be made publicly available to the research community on open digital platforms upon completion (a link to a sample dataset is available upon request).

Keywords: CISPD (Corpus of Italian and Serbian Public Discourse), Italian language, Serbian language, comparative corpus, public discourse, spoken discourse, written discourse

Works are licensed under the terms and conditions of the Creative Commons Attribution 4.0 International License.

<https://creativecommons.org/licenses/by/4.0/>



Croatian Tour-Guiding Corpus for AI-Augmented Japanese Language and Culture Learning

Irena Srdanović¹[\[0000-0003-1281-176X\]](#), Benedikt Perak²[\[0000-0003-4177-5307\]](#),
Dragana Špica¹[\[0000-0003-1008-2700\]](#)

¹ University of Jurja Dobrile in Pula, Faculty of Humanities, Croatia,
{irena.srdanovic|dragana.spica}@unipu.hr

² University of Rijeka, Faculty of Humanities and Social Science, Croatia,
bperak@uniri.hr

Abstract

This paper presents the development of a specialized Croatian-language corpus of tour-guiding discourse aimed at supporting the education and early career development of future guides working with Japanese tourists. The corpus is based on multilingual audio and video materials collected through practical training activities designed to support students' early career development in tourism. Students, university teachers, licensed tour guides, translators, and interpreters participated in simulated on-site and virtual tours conducted in Croatian, English, and Japanese. The resulting materials were transcribed in multiple languages using AI-assisted technologies, and then organized and analyzed using Sketch Engine (Kilgarriff et al., 2004).

Building on the Japanese corpus previously developed from the same dataset, this study focuses on Croatian transcripts from tours of museums, cities, heritage sites, attractions, virtual tours, storytelling activities, and professional guidance sessions. The analysis examines tourism-specific vocabulary and discourse relating to Croatian destinations, local food, tangible and intangible cultural heritage, historical explanation, guiding practices, and culturally relevant advice for future guides of Japanese-speaking visitors (Srdanović et al., 2013).

The corpus is conceived as part of VIRAI – Language for Real and Virtual Worlds, a project developing a foreign-language learning model for tourism through corpora, artificial intelligence, and extended reality. Validated lexical and contextual data will contribute to the Japanese–Croatian WaKu Learner's Dictionary and to the development of a searchable Croatian–Japanese parallel corpus, with future integration into the

SMIRAI virtual language and cultural learning environment (Srdanović et al., 2026). Given the linguistic proximity of South Slavic languages, the proposed approach may also provide a transferable model for developing comparable resources for Serbian and other South Slavic languages and for integrating local linguistic and cultural knowledge into foreign-language education. Together, these resources are intended to support corpus-grounded, AI-augmented, and context-sensitive learning of Japanese language, culture, and professional tour-guiding communication.

Keywords: Croatian tour-guiding corpus, Japanese for tourism, specialized vocabulary, AI-augmented learning, Sketch Engine, VIRAI, SMIRAI

Acknowledgements

This work was funded by the EU NextGeneration under the Juraj Dobrila University of Pula institutional research project “From Dynamic Vocabulary to Virtual Interaction: Development of a Foreign Language Learning Model in Tourism Using Corpora, Artificial Intelligence, and Extended Reality (VIRAI – Language for Real and Virtual Worlds) (project number: IIP_UNIPU₀10138, leader: Prof. Irena Srdanović).

References

- Adam Kilgarriř, Pavel Rychlý, Pavel Smrž, and David Tugwell. 2004. *The Sketch Engine*. In *Proceedings of the 11th EURALEX International Congress*, pages 105–116, Lorient, France.
- Irena Srdanović, Aida Smajić, and Martina Rovis. 2026. *SMIRAI: A Prototype of a Virtual Cultural and Language Learning Tutor*. *Acta Linguistica Asiatica*, 16(2):263–285.
- Irena Srdanović, Vít Suchomel, Toshinobu Ogiso, and Adam Kilgarriř. 2013. *Japanese Language Lexical and Grammatical Profiling Using the Web Corpus JpTenTen*. In *Proceedings of the 3rd Japanese Corpus Linguistics Workshop*, pages 229–238, Tokyo. NINJAL, Department of Corpus Studies/Center for Corpus Development.

Hrvatski korpus turističkog vođenja za učenje japanskog jezika i kulture uz podršku veštačke inteligencije

Irena Srdanović, Benedikt Perak, Dragana Špica

Sažetak

Ovaj rad predstavlja razvoj specijalizovanog korpusa diskursa turističkog vođenja na hrvatskom jeziku, namenjenog podršci obrazovanju i ranom razvoju karijere budućih vodiča koji će raditi sa japanskim turistima. Korpus se zasniva na višejezičnim audio i video materijalima prikupljenim kroz aktivnosti stručne prakse usmerene na podršku ranom razvoju karijere studenata u turizmu. Studenti, univerzitetski nastavnici, licencirani turistički vodiči, prevodioci i usmeni prevodioci učestvovali su u simuliranim terenskim i virtuelnim turama koje su se odvijale na hrvatskom, engleskom i japanskom jeziku. Dobijeni materijali transkribovani su na više jezika uz pomoć tehnologija zasnovanih na veštačkoj inteligenciji, a zatim organizovani i analizirani pomoću alata Sketch Engine.

Nadovezujući se na korpus na japanskom jeziku, prethodno razvijen na osnovu istog skupa podataka, ovo istraživanje usredsređuje se na transkripte na hrvatskom jeziku koji obuhvataju ture po muzejima, gradovima, lokalitetima kulturne baštine i turističkim atrakcijama, virtuelne ture, aktivnosti pripovedanja (storytelling) i sesije profesionalnog usmeravanja. Analizom se ispituju vokabular i diskurs specifični za turizam, koji se odnose na hrvatske destinacije, lokalnu gastronomiju, materijalnu i nematerijalnu kulturnu baštinu, tumačenje istorijskih sadržaja, prakse turističkog vođenja i kulturološki relevantne smernice za buduće vodiče japanskih turista.

Korpus zamišljen je kao deo projekta VIRAI – Language for Real and Virtual Worlds, u okviru kojeg se razvija model učenja stranih jezika za potrebe turizma primenom korpusa, veštačke inteligencije i proširene stvarnosti. Validirani leksički i kontekstualni podaci doprineće razvoju japansko-hrvatskog učeničkog rečnika WaKu i pretraživog hrvatsko-japanskog paralelnog korpusa, uz planiranu buduću integraciju u virtuelno okruženje SMIRAI za učenje jezika i kulture. S obzirom na jezičku bliskost južnoslovenskih jezika, predloženi pristup može poslužiti i kao prenosiv model za razvoj uporedivih resursa za srpski i druge južnoslovenske jezike, kao i za uključivanje lokalnog jezičkog i kulturnog znanja u nastavu stranih jezika. Zajedno, ovi resursi treba da omoguće učenje japanskog jezika, kulture i profesionalne komunikacije u turističkom vođenju koje je utemel-

jeno na korpusnim podacima, potpomognuto veštačkom inteligencijom i prilagođeno konkretnom kontekstu.

Ključne reči: hrvatski korpus turističkog vođenja, japanski jezik za turizam, specijalizovani vokabular, učenje uz podršku veštačke inteligencije, Sketch Engine, VIRAI, SMIRAI

Works are licensed under the terms and conditions of the Creative Commons Attribution 4.0 International License.

<https://creativecommons.org/licenses/by/4.0/>





Abstracts of Papers

TESLA

ELEXIS-WSD 2.0: Expanding the Lexicon–Corpus Interface with New Languages and Annotation Layers

Jaka Čibej¹[0000-0002-3037-6848], Simon Krek¹[0000-0001-8965-6863], Carole Tiberius²[0000-0002-7860-5427], Irina Lobzhanidze³[0000-0002-3040-1847],
Verginica Barbu Mititelu⁴[0000-0003-1945-2587], Jelena Kallas⁵[0000-0001-8614-0140], Kertu Saul⁵[0009-0002-0964-0015], Kadri Muischnek⁶[0000-0002-3920-6618], Karoliina Jõgi⁶, Ranka Stanković⁷[0000-0001-5123-6273], Cvetana Krstev⁷[0000-0003-3328-9392],
Aleksandra Marković⁷[0000-0003-2985-3042], Milica Ikonić Nešić⁷[0000-0002-0835-8889], Ana Ostroški Anić⁸[0000-0001-9999-0750],
Bartłomiej Alberski⁹[0000-0002-9099-6627], Vladimir Cvetkoski¹⁰[0000-0002-8570-1062], Voula Giouli¹¹[0000-0002-1501-9754],
Rusudan Makhachashvili¹²[0000-0002-4806-6434], Olha Kanishcheva¹³[0000-0002-9035-1765]

¹ University of Ljubljana, Slovenia, {jaka.cibej|simon.krek}@ff.uni-lj.si

² Dutch Language Institute, The Netherlands, carole.tiberius@ivdnt.org

³ Iliia State University, Georgia, irina_lobzhanidze@iliauni.edu.ge

⁴ Research Institute for Artificial Intelligence, Romania, vergi@racai.ro

⁵ Institute of the Estonian Language, Estonia, {jelena.kallas|kertu.saul}@eki.ee

⁶ University of Tartu, Estonia, {kadri.muischnek|karoliina.jogi}@ut.ee

⁷ Language Resources and Technologies Society – JeRTeh, Serbia, {cvetana|ranka|milicain}@jerteh.rs, aleksandra.markovic@isj.sanu.ac.rs

⁸ Institute for the Croatian Language, Croatia, aostrosk@ihjj.hr

⁹ Wrocław University of Science and Technology, Poland,
bartlomiej.alberski@pwr.edu.pl

¹⁰ Ss. Cyril and Methodius University in Skopje, North Macedonia,
cvetkoski@ff.ukim.edu.mk

¹¹ Aristotle University of Thessaloniki, Greece, pgiouli@del.auth.gr

¹² Borys Grinchenko Kyiv Metropolitan University, Ukraine,
r.makhachashvili@kubg.edu.ua

¹³ Heidelberg University, Germany, kanichshevaolga@gmail.com

Abstract

The ELEXIS-WSD Parallel Sense-Annotated Corpus is a multilingual resource designed to support word sense disambiguation, lexical-semantic research, and cross-lingual studies by linking corpus occurrences directly to lexical resources through manual sense annotation. Within the UniDive COST Action, WG2 (Task 2.2) focused on extending and enriching this resource, resulting in the release of ELEXIS-WSD 2.0 in April 2026 (Čibej et al., 2026). The new version significantly expands the corpus by introducing eight additional languages: Romanian, Greek, Ukrainian, Polish, Macedonian, Georgian, Croatian, and Serbian, while further enriching existing subcorpora for Slovene, Estonian, and Dutch.

The extension process involved multiple annotation layers. Newly added language subcorpora were compiled through automatic translation of 2,024 English source sentences followed by manual correction and validation. The corrected texts were tokenized, lemmatized, morphosyntactically annotated, and syntactically parsed, with manual corrections performed where resources permitted. Additional semantic layers include the annotation of multiword expressions according to the PARSEME 2.0 guidelines and named entities linked to Wikidata. The Serbian team contributed with named entity annotation and linking, as well as the first sense-annotated subcorpus through a semi-automatic annotation workflow.

As a result, ELEXIS-WSD 2.0 now comprises approximately 38,000 sentences and 620,000 tokens, representing an increase of about 80% compared to the pre-UniDive version. The corpus is openly available through the CLARIN.SI repository and provides a valuable infrastructure for multilingual lexical-semantic research. Future work will focus on completing remaining annotation layers, harmonizing annotations across languages, expanding sense inventories, and exploring the use of large language models for automatic sense generation and linking.

Keywords: word sense disambiguation, parallel corpus, semantic annotation, named entities, multiword expressions, lexical resources, ELEXIS-WSD

ELEXIS-WSD 2.0: Proširenje interfejsa između leksikona i korpusa novim jezicima i slojevima anotacije

Jaka Čibej, Simon Krek, Carole Tiberius, Irina Lobzhanidze, Verginica Barbu Mititelu, Jelena Kallas, Kertu Saul, Kadri Muischnek, Karoliina Jõgi, Ranka Stanković, Cvetana Krstev, Aleksandra Marković, Milica Ikonić Nešić, Ana Ostroški Anić, Bartłomiej Alberski, Vladimir Cvetkoski, Voula Giouli, Rusudan Makhachashvili, Olha Kanishcheva

Sažetak

Paralelni korpus sa anotiranim značenjima ELEXIS-WSD predstavlja višejezički resurs namenjen podršci zadacima razrešavanja značenja reči, leksičko-semantičkim istraživanjima i međujezičkim analizama, povezivanjem korpusnih primera sa leksičkim resursima putem ručne anotacije značenja. U okviru COST akcije UniDive, Radna grupa 2 (WG2) bila je posvećena proširenju i unapređenju ovog resursa, što je rezultovalo objavljivanjem verzije ELEXIS-WSD 2.0 u aprilu 2026. godine. Nova verzija značajno proširuje korpus uvođenjem osam novih jezika: rumunskog, grčkog, ukrajinskog, poljskog, makedonskog, gruzijskog, hrvatskog i srpskog, dok su postojeći potkorpusi za slovenački, estonski i holandski dodatno obogaćeni novim slojevima anotacije.

Proces proširenja obuhvatio je više anotacionih slojeva. Novi jezički potkorpusi nastali su automatskim prevodenjem 2.024 izvorne engleske rečenice, nakon čega su usledile ručna korekcija i validacija prevoda. Tekstovi su zatim tokenizovani, lematizovani, morfosintaksički anotirani i sintaksički parsirani, uz ručne ispravke gde su raspoloživi resursi to omogućavali. Dodatni semantički slojevi uključuju anotaciju višečlanih izraza prema smernicama PARSEME 2.0 i anotaciju imenovanih entiteta povezanih sa bazom Wikidata. Srpski tim dao je značajan doprinos kroz anotaciju i povezivanje imenovanih entiteta, kao i izradom prvog potkorpusa sa anotiranim značenjima primenom poluautomatskog postupka anotacije.

Kao rezultat, ELEXIS-WSD 2.0 sada obuhvata približno 38.000 rečenica i 620.000 tokena, što predstavlja povećanje od oko 80% u odnosu na verziju korpusa pre početka projekta UniDive. Korpus je javno dostupan putem repozitorijuma CLARIN.SI i predstavlja značajnu infrastrukturu za višejezička leksičko-semantička istraživanja. Budući rad biće usmeren na završetak preostalih anotacionih slojeva, harmonizaciju anotacija između

jezika, proširenje inventara značenja i istraživanje mogućnosti primene velikih jezičkih modela za automatsko generisanje i povezivanje značenja.

Ključne reči: razrešavanje značenja reči, paralelni korpus, semantička anotacija, imenovani entiteti, višečlani izrazi, leksički resursi, ELEXIS-WSD

Repozitorijum simila savremenog srpskog jezika – izgradnja i evaluacija mašinski potpomognute anotacije

Milica Dinić Marinković¹[\[0000-0002-2641-8806\]](#),
Cvetana Krstev²[\[0000-0003-3328-9392\]](#),
Aleksandra Marković³[\[0000-0003-2985-3042\]](#),
Ranka Stanković⁴[\[0000-0001-5123-6273\]](#)

¹Univerzitet u Beogradu, Filološki fakultet, Srbija,
milica.dinic.marinkovic@fil.bg.ac.rs

²Društvo za jezičke resurse i tehnologije – JeRTeh, Srbija, cvetana@jerteh.rs

³Institut za srpski jezik Srpske akademije nauka i umetnosti, Srbija,
aleksandra.markovic@isj.sanu.ac.rs

⁴Univerzitet u Beogradu, Rudarsko-geološki fakultet, Srbija,
ranka.stankovic@rgf.bg.ac.rs

Apstrakt

Cilj rada je evaluacija metodologije izgradnje *Repozitorijuma simila savremenog srpskog jezika* (SSR), razvijenog u okviru šireg istraživačkog projekta konstruisanja prvog korpusno zasnovanog jezičkog resursa za sistematsko proučavanje simila u srpskom. SSR objedinjuje korpusnu ekstrakciju, semantičku anotaciju, lingvističku validaciju i strukturirano organizovanje podataka u jedinstven, otvoren i proširiv istraživački resurs.

U radu se predstavljaju rezultati svih faza obrade – od korpusne ekstrakcije kandidata i filtriranja zasnovanog na pravilima do automatske semantičke anotacije i njene ručne validacije. Repoziotorijum se zasniva na više komplementarnih korpusa savremenog srpskog jezika. Za automatsku obradu primenjen je hibridni neuro-simbolički pristup. U ekstrakciji se kombinuju korpusni upiti i lokalne gramatike realizovane kao konačni automati u sistemu Unitex, koje koriste elektronske rečnike srpskog jezika, dok su za automatsko prepoznavanje i anotaciju kandidata izdvojenih u prethodnim koracima testirani veliki jezički modeli (dalje VJM) GPT-5.2 i Gemini.

Budući zamišljen kao dinamički resurs namenjen lingvističkim, leksikografskim i istraživanjima u okviru obrade prirodnog jezika, ali i kao osnova za dalji razvoj metoda za automatsku obradu figurativnog jezika u srpskom, u fokusu ovog rada su anotacioni model, rezultati lingvističke eva-

luacije mašinski potpomognute anotacije i implikacije dobijenih rezultata za unapređenje razvijenog sistema.

Evaluacija pokazuje da kombinacija simboličkih metoda i VJM omogućava visok stepen preciznosti u identifikaciji simila, uz zadovoljavajuću saglasnost sa stručnim anotatorima. Istovremeno, analiza upućuje da uspešna identifikacija i anotacija simila zahteva uvažavanje složenih odnosa između sintaksičke strukture, semantičke interpretacije i implicitno realizovanih elemenata poređenja, što predstavlja dodatne izazove za automatsku anotaciju. Dobijeni rezultati potvrđuju značaj hibridnog pristupa koji objedinjuje simboličke pristupe i upotrebu VJM sa lingvističkom ekspertizom ugrađenom u anotacionu shemu, evaluaciju i iterativno unapređenje modela.

Ključne reči: Simili, repozitorijum simila, srpski jezik, korpusna lingvistika, semantička anotacija, leksički resursi, figurativni jezik, Unix, veliki jezički modeli.

The Repository of Similes in Contemporary Serbian: Construction and Evaluation of Machine-Assisted Annotation

Milica Dinić Marinković, Cvetana Krstev, Aleksandra Marković, Ranka Stanković

Summary

This paper evaluates the methodology developed for constructing the *Similes Repository for Contemporary Serbian* (SSR), created within a broader research project to build the first corpus-based language resource for the systematic study of similes in Serbian. SSR integrates corpus extraction, semantic annotation, linguistic validation, and structured data organization into a unified, open, and extensible resource.

The paper presents the complete processing pipeline, from corpus-based candidate extraction and rule-based filtering to automatic semantic annotation and manual validation. SSR is based on several complementary corpora of Contemporary Serbian. Automatic processing relies on a hybrid neuro-symbolic approach that combines corpus queries and local grammars implemented as FSTs in Unitex, supported by electronic dictionaries of Serbian, with GPT-5.2 and Gemini LLMs for automatic simile identification and semantic annotation.

Conceived as a dynamic resource for linguistic, lexicographic, and NLP research, as well as a foundation for the further development of automatic figurative language processing in Serbian, the paper focuses on the annotation model, linguistic evaluation of machine-assisted annotation, and implications for further improving the proposed framework.

The evaluation demonstrates that combining symbolic methods with LLMs achieves high precision in simile identification while maintaining satisfactory agreement with expert annotators. At the same time, the analysis shows that successful simile identification and annotation requires accounting for complex interactions between syntactic structure, semantic interpretation, and implicit comparison components. The results confirm the effectiveness of a hybrid methodology that integrates symbolic methods, LLMs, and linguistic expertise throughout annotation design, evaluation, and iterative refinement of the framework.

Keywords: Similes, Simile Repository, Serbian Language, Corpus Linguistics, Semantic Annotation, Lexical Resources, Figurative Language, Unitex, Large Language Models.

srWOZ: Towards Task-Oriented Dialogue Modelling in Serbian

Anđelka Zečević¹[0000-0002-5656-007X],
Jovana Rađenović^{2,5}[0000-0002-2707-3870], Milica Gašić^{1,3}[0000-0003-0318-9147],
Songbo Hu⁴[0000-0003-1695-459X], Ranka Stanković²[0000-0001-5123-6273],
Mihailo Škorić²[0000-0003-4811-8692]

¹ Serbian Academy of Sciences and Arts, Mathematical Institute, Serbia,
andjelkaz@mi.sanu.ac.rs

² University of Belgrade, Faculty of Mining and Geology, Serbia,
{jovana.radjenovic|ranka.stankovic|mihailo.skoric}@rgf.bg.ac.rs

³ Heinrich Heine University Düsseldorf, Germany, gasic@hhu.de

⁴ University of Cambridge, Language Technology Lab, United Kingdom,
sh2091@cam.ac.uk

⁵ University of Belgrade, Faculty of Mathematics, Serbia

Abstract

The rapid adoption of large language models and agentic frameworks has increased the demand for high-quality conversational datasets for the development and evaluation of dialogue systems. Among these, task-oriented dialogue systems have emerged as a particularly important class of conversational systems, assisting users in accomplishing specific goals through natural language interaction. Despite their growing popularity and practical importance, research and development of task-oriented dialogue systems for Serbian remain limited due to the scarcity of adequate language resources and annotated dialogue datasets.

In this paper, we present *srWOZ*, a task-oriented multi-turn conversational dataset for Serbian. Its creation is inspired by the existing *Multi³WOZ* dataset, a multilingual multi-domain multi-parallel dataset of culturally adapted task-oriented dialogues. In order to create *srWOZ*, we followed the original *Multi³WOZ* language adaptation methodology. First, we adapted the underlying ontology to the Serbian cultural and institutional context, covering seven domains: attractions, hospitals, police services, hotels, restaurants, taxis, and public transportation, while also performing slot-value normalization. Second, we adapted outline-based

dialogue generation component by adapting dialogue schemas and dialogue writing modules. Finally, we collected dialogues from native Serbian speakers with a bachelor degree in Serbian language using a bottom-up outline-based method. This process resulted in an initial dataset comprising 812 dialogues.

By addressing the lack of conversational resources for Serbian, *sr-WOZ* contributes to the broader effort of extending dialogue technologies to low-resource linguistic settings by highlighting the process and the difficulty and best practices of such an endeavour.

Keywords: dialogue system, task-oriented conversations, cultural adaptation

srWOZ: ka modelovanju dijaloga usmerenih na zadatak u srpskom jeziku

Sažetak

Nadasve brzo odomaćivanje velikih jezičkih modela i agentskih okvira otvorilo je potrebu za kreiranjem visokokvalitetnih skupova podataka koji se mogu koristiti za razvoj i evaluaciju sistema za dijalog. Među njima se posebno izdvajaju *sistemi za dijalog usmereni na zadatak* (engl. *task-oriented dialogue systems*) kao važna klasa konverzacijskih sistema koja pomaže korisnicima da ostvare konkretne zadatke i ciljeve kroz interakcije na prirodnom jeziku. Uprkos njihovoj rastućoj popularnosti i praktičnom značaju, istraživanje i razvoj ovakvih sistema za srpski jezik i dalje su ograničeni zbog nedostatka odgovarajućih jezičkih resursa i anotiranih skupova dijaloških podataka.

U ovom radu predstavljamo *srWOZ*, skup podataka za srpski jezik koji sadrži *višekružne* dijaloge usmerene na zadatak. Njegovo kreiranje inspirisano je postojećim skupom podataka *Multi³WOZ*, višejezičnim, višedomenskim i višeparalelnim skupom kulturno prilagođenih dijaloga orijentisanih na zadatak. Za potrebe razvoja skupa podataka *srWOZ*, pratili smo originalnu metodologiju jezičke adaptacije primenjenu u okviru projekta *Multi³WOZ*. Najpre smo prilagodili osnovnu ontologiju srpskom kulturnom i institucionalnom kontekstu, obuhvativši sedam domena: turističke atrakcije, bolnice, policijske službe, hotele, restorane, taksi prevoz i javni prevoz, uz normalizaciju vrednosti slotova. Zatim smo prilagodili komponentu za generisanje dijaloga zasnovanu na shemama (engl. *outline-based*), adaptacijom odgovarajućih dijaloških shema i modula za zapisivanje dijaloga. Konačno, prikupili smo dijaloge od strane izvornih govornika srpskog jezika primenom *odozdo-nagore* pristupa zasnovanog na shemama (engl. *bottom-up outline-based method*). Ovim procesom formiran je inicijalni skup podataka sa 817 dijaloga.

Nadomešćujući problem nedostatka resursa za modelovanje dijaloga na srpskom jeziku, *srWOZ* doprinosi širim naporima jezičke zajednice usmerenim ka proširenju dijaloških tehnologija na jezička okruženja sa ograničenim resursima poput srpskog jezičkog okruženja.

Ključne reči: sistemi za dijalog, dijalog orijentisan na zadatak, kulturna adaptacija

Речник савременог српског језика: од дигиталне лексикографске базе до отворене језичке инфраструктуре

Ранка Станковић¹[\[0000-0001-5123-6273\]](#), Михаило
Шкорић¹[\[0000-0003-4811-8692\]](#), Рада Стијовић¹

¹ Друштво за језичке ресурсе и технологије – ЈеРТех, Србија
{ranka|mihailo}@jerteh.rs, stijovicr@gmail.com

Апстракт

Речник савременог српског језика (РССЈ) представља лексикографски подухват чији је циљ изградња савременог речника српског језика заснованог на корпусним подацима, савременим лексикографским принципима и језичким технологијама. За разлику од традиционалних речника, РССЈ је замишљен као интегрисана дигитална инфраструктура која обједињује речничку базу података, лексикографско окружење, корпусне ресурсе, програмске интерфејсе, веб и мобилне апликације. Рад представља архитектуру система и модел података имплементиран у систему PostgreSQL, као и интеграцију са корпусом СрпКор2021+ на платформи NoSketch Engine за претрагу корпуса. Посебно се разматра концепт програмског интерфејса REST API, који омогућава да речничким подацима приступају и људи и машине, чиме се речник повезује и са другим језичким ресурсима, попут лексичке базе Лексимирика и Српског ворднета. Пажња је посвећена и веб и мобилним апликацијама, механизмима аутентификације и управљања приступом, као и могућностима интеграције са системима заснованим на вештачкој интелигенцији. Рад показује како се традиционална лексикографија може трансформисати кроз савремену дигиталну инфраструктуру способну да подржи истраживачки рад, образовање, индустријске примене и развој језичких технологија за српски језик.

Кључне речи: дигитална лексикографија, речничке базе података, API, веб-апликације, мобилне апликације, језичке технологије, српски језик

The Dictionary of Contemporary Serbian Language: From a Digital Lexicographic Resource to an Open Language Infrastructure

Ranka Stanković, Mihailo Škorić, Rada Stijović

Summary

The Dictionary of Contemporary Serbian Language (RSSJ) is a lexicographic initiative aimed at developing a modern Serbian dictionary based on corpus data, contemporary lexicographic principles, and language technologies. Unlike traditional dictionaries, the RSSJ is conceived as an integrated digital infrastructure that combines a lexical database, a lexicographic working environment, corpus resources, application programming interfaces, and web and mobile applications. This paper presents the overall system architecture and the underlying data model, aligned with the DMLex standard and implemented in PostgreSQL, as well as its integration with the SrpKor corpora and the NoSketch Engine platform. Special attention is given to the concept of a REST API that enables access to lexical data for both human users and software applications, thereby connecting the dictionary with other language resources such as the Leximirka lexical database and the Serbian WordNet. The paper also addresses the web and mobile applications, authentication and access-control mechanisms, and possibilities for integration with AI systems. It demonstrates how traditional lexicography can be transformed into a modern digital infrastructure capable of supporting research, education, industrial applications, and the development of language technologies for Serbian.

Keywords: digital lexicography, lexical databases, API, web applications, mobile applications, language technologies, Serbian language.

Од (слике) текста ка графу знања: дигитализација и семантичко обогаћивање енциклопедије Свезнање

Јована Рађеновић^{1,2}[0000-0002-2707-3870],
Биљана Рујевић¹[0000-0002-9103-3902],
Александра Томашевић¹[0000-0003-1583-7447],
Оливера Китановић¹[0000-0002-7571-2729],
Ранка Станковић¹[0000-0001-5123-6273]

¹ Универзитет у Београду, Рударско-геолошки факултет, Србија,
{jovana.radjenovic|biljana.rujevic|aleksandra.tomasevic|olivera.kitanovic
|ranka.stankovic}@rgf.bg.ac.rs

² Универзитет у Београду, Математички факултет, Србија

Апстракт

Рад представља поступак дигитализације, структурног обогаћивања и изградње графа знања заснованог на историјској српској енциклопедији „Свезнање“ из 1937. године. Процес је обухватио скенирање, оптичко препознавање текста, ручну корекцију грешака, структурно уређивање и конверзију у формат TEI-XML. Посебна пажња посвећена је очувању изворног правописа и структуре текста, као и дефинисању прецизних упутстава за сараднике који су учествовали у корекцији и анотацији материјала. У раду су описани најчешћи проблеми дигитализације, укључујући мешање ћирилице и латинице, погрешно препознавање карактера и структурне грешке у тексту.

Након корекције текста примењене су методе препознавања именованих ентитета (NER), повезивања ентитета са базом знања Википодаци (NEL), као и екстракције релација између ентитета. Енциклопедијски чланци су додатно класификовани према домену и семантичком типу, чиме је омогућена изградња семантички обогаћеног графа знања. Добијени ресурс представља значајан допринос дигиталној хуманистици и развоју језичких технологија за српски језик, омогућавајући напредну семантичку претрагу, анализу културног наслеђа и интеграцију са екосистемом повезаних отворених података.

Додатно, применом великих језичких модела над препричаним и сажетим верзијама одредница генерисан је и скуп парова питања и

одговора заснован на садржају енциклопедије, намењен даљим применама у системима за одговарање на питања и приступима заснованим на генерисању потпомогнутом претрагом (RAG).

Кључне речи: српски језик, дигитализација, графови знања, обрада природног језика, препознавање именованих ентитета, повезивање именованих ентитета, генерисање питања и одговора

From Text (Image) to a Knowledge Graph: Digitization and Semantic Enrichment of the Sveznanje Encyclopedia

Jovana Rađenović, Biljana Rujević, Aleksandra Tomašević, Olivera Kitanović, Ranka Stanković

Summary

This paper presents the process of digitization, structural enrichment, and knowledge graph construction based on the historical Serbian encyclopedia „Свезнање“, first published in 1937. The workflow included scanning, OCR-based text recognition, manual error correction, structural editing, and conversion into the TEI-XML format. Special attention was given to preserving the original orthography and document structure, as well as to defining precise annotation and correction guidelines for collaborators involved in the project. The paper discusses the most common OCR-related issues, including script mixing between Cyrillic and Latin alphabets, character recognition errors, and structural inconsistencies in the text.

Following text correction, Named Entity Recognition (NER), Named Entity Linking (NEL) to Wikidata, and relation extraction techniques were applied. Encyclopedia entries were additionally classified according to their domain and semantic type, enabling the construction of a semantically enriched knowledge graph. The resulting resource represents an important contribution to digital humanities and Serbian language technologies, providing support for advanced semantic search, cultural heritage analysis, and integration into the Linked Open Data ecosystem.

In addition, large language models were used to generate a complementary set of question-answer pairs from the encyclopedia entries, intended for further use in question-answering and retrieval-augmented generation (RAG) systems.

Keywords: Serbian language, digitization, knowledge graphs, natural language processing, named entity recognition, named entity linking, question-answer generation

Hierarchical Recognition of Moral Foundations in Serbian Literary Corpus *SrpELTeC*

Milena Šošić¹[\[0000-0002-9949-9686\]](#), Ranka Stanković²[\[0000-0001-5123-6273\]](#)

¹ Language Resources and Technologies Society – JeRTeh, Serbia,
milena.sosic@jerteh.rs

²University of Belgrade, Faculty of Mining and Geology, Serbia,
ranka.stankovic@rgf.bg.ac.rs

Abstract

This paper presents a novel approach for the hierarchical moral analysis based on the Moral Foundations Theory (MFT) in Serbian literary corpus *SrpELTeC*. The proposed method automatically identifies and quantifies five moral foundations together with their corresponding virtue and vice categories – care/harm, fairness/cheating, loyalty/betrayal, authority/subversion and purity/degradation, within different segments of Serbian novels. The computation was performed using a developed lexicon of moral words *MFD.SR*, adapted to the linguistic and cultural characteristics of Serbian language. A key methodological contribution of this work is the recognition of moral signals across multiple levels of a hierarchically structured corpus, which enables comparable moral analysis of texts of varying sizes and characteristics. In this way, the method produces both macro and micro level representations of moral dynamics within literary narratives, that allows fine-grained tracking of shifts in moral values across different corpus segments. Preliminary results indicate that Serbian novels first published in 1840-1920. time period, exhibit distinct and measurable moral signatures across authors and literary movements, which offers new insights into the ideological and cultural structure of Serbian literature from this period.

Keywords: moral foundations, sentiment, lexicon, literary corpus, Serbian language

Hijerarhijsko prepoznavanje moralnih vrednosti u književnom korpusu *SrpELTeC*

Milena Šošić, Ranka Stanković

Sažetak

Ovaj rad predstavlja novi pristup za hijerarhijsku analizu morala zasnovanoj na Teoriji moralnih vrednosti (*Moral Foundations Theory, MFT*) u okviru korpusa srpske književnosti *SrpELTeC*. Predloženi metod automatski identifikuje i kvantifikuje pet osnovnih moralnih vrednosti zajedno sa njima odgovarajućim kategorijama sentimenta – briga/šteta (eng. care/harm), pravičnost/prevara (eng. fairness/cheating), lojalnost/izdaja (eng. loyalty/betrayal), autoritet/subverzija (eng. authority/subversion) i čistota/degradacija (eng. purity/degradation), u okviru različitih segmenata srpskih romana. Izračunavanje je izvršeno korišćenjem razvijenog leksikona moralnih reči *MFD.SR*, prilagođenog lingvističkim i kulturološkim karakteristikama srpskog jezika. Ključni metodološki doprinos ovog rada predstavlja prepoznavanje moralnih signala na više nivoa hijerarhijski strukturisanog korpusa, što omogućava uporedivu analizu tekstova različitih veličina i karakteristika. Na ovaj način, predloženi metod generiše reprezentacije moralne dinamike na makro i mikro nivou u okviru književnih narativa, čime se omogućava granularno praćenje promena moralnih vrednosti kroz različite segmente korpusa. Preliminarni rezultati ukazuju da srpski romani prvi put objavljeni u 1840-1920. periodu, ispoljavaju prepoznatljive i merljive moralne obrasce među autorima i književnim pravcima, čime se otvaraju novi uvidi u ideološku i kulturološku strukturu srpske književnosti ovog vremenskog perioda.

Ključne reči: moralne vrednosti, sentiment, leksikon, književni korpus, srpski jezik

Named Entity Recognition and Relation Extraction in Serbian Football Text

Jelena Lazarević¹[0009-0004-4481-2729], Ranka Stanković²[0000-0001-5123-6273],
Olivera Kitanović²[0000-0002-7571-2729]

¹ Language Resources and Technologies Society – JeRTeh, Serbia,
jelazarevic1@gmail.com

²University of Belgrade, Faculty of Mining and Geology, Serbia,
{ranka.stankovic|olivera.kitanovic}@rgf.bg.ac.rs

Abstract

The language of football represents a highly dynamic and semantically rich domain for Natural Language Processing (NLP), particularly for tasks such as: Named Entity Recognition (NER), Named Entity Linking (NEL), and Relation Extraction (RE). The discourse of football contains numerous entities and complex semantic relations including: players, clubs, coaches, competitions, transfers, injuries, and match events. This paper focuses on information extraction from the Serbian football corpus *SrFudKo*, originally developed for contrastive discourse analysis and corpus-based linguistic research. Previous work resulted in the construction of Serbian and Spanish football corpora, while semantic annotation tasks have not yet been systematically explored, making this precious linguistic resource as a basis for the knowledge graph construction as well as semantic search. The goal of this research is to enrich the Serbian football corpus with semantic annotations and structured relational information. The paper investigates the methods for recognizing football-specific named entities, linking them to external knowledge bases, such as Wikidata, and extracting semantic relations between entities. Paying special attention to the football-specific relation types, we included the relations between players, clubs, coaches, competitions, and events, as well as their organization within a domain-oriented ontology. In addition to presenting the corpus and the proposed ontology, the paper provides quantitative analysis and illustrative examples of extracted entities and their relations, based on the Serbian football-themed articles. The presented approach contributes to the development of semantically enriched football resources, suitable for knowledge graph population, and domain-specific NLP applications.

Keywords: football, corpus, NER, NEL, relation extraction, Serbian

Препознавање именованих ентитета и екстракција релација у српским фудбалским текстовима

Јелена Лазаревић, Ранка Станковић, Оливера Китановић

Сажетак

Језик фудбала представља веома динамичну и семантички богату област за обраду природног језика, посебно за задатке као што су препознавање именованих ентитета (Named Entity Recognition – NER), повезивање именованих ентитета (Named Entity Linking – NEL) и екстракција релација (Relation Extraction – RE). Фудбалски дискурс садржи бројне ентитете и сложене семантичке односе који укључују играче, клубове, тренере, такмичења, трансфере, повреде и догађаје са утакмица. Овај рад је усмерен на екстракцију информација из српског фудбалског корпуса *SrFudKo*, који је првобитно развијен за контрастивну анализу дискурса и корпусна лингвистичка истраживања. Претходна истраживања су резултовала формирањем српског и шпанског фудбалског корпуса, а у наставку истраживања постављени су задаци семантичког аотирања, те *SrFudKo* драгоцени лингвистички ресурс чини основу за изградњу графова знања и семантичку претрагу. Циљ овог истраживања јесте обogaћивање српског фудбалског корпуса семантичким аотацијама и структурираним релационим информацијама. У раду се истражују методе за препознавање фудбалски специфичних именованих ентитета, њихово повезивање са спољним базама знања као што је Википодаци, као и екстракција семантичких релација између ентитета. Посебна пажња посвећена је типовима релација карактеристичним за фудбалски домен, укључујући односе између играча, клубова, тренера, такмичења и догађаја, као и њиховој организацији у оквиру доменски оријентисане онтологије. Поред представљања корпуса и предложене онтологије, рад доноси квантитативне лингвистичке анализе и илустративне примере екстрахованих ентитета и релација из српских фудбалских текстова. Предложени приступ доприноси развоју семантички обогaћених фудбалских ресурса погодних за корпусну лингвистику, попуњавање графова знања, екстракцију информација и доменски специфичне NLP примене.

Кључне речи: фудбал, корпус, NER, NEL, екстракција релација, српски језик

Lexicon-Based Exploration of Emotion Dynamics in TEI Novels: Case Study on *SrpELTeC* Corpus

Ranka Stanković¹[\[0000-0001-5123-6273\]](#), Milena Šošić²[\[0000-0002-9949-9686\]](#)

¹University of Belgrade, Faculty of Mining and Geology, Serbia,
ranka.stankovic@rgf.bg.ac.rs

² Language Resources and Technologies Society – JeRTeh, Serbia,
milena.sosic@jerteh.rs

Abstract

This paper presents a hierarchical analysis of emotional dynamics in the *SrpELTeC* corpus of Serbian novels from the period 1850–1920s. The study combines TEI-encoded texts with a Serbian-tailored emotion lexicon *EmoLex.SR*, based on Plutchik’s model of basic emotions. Emotional signals from the corpus were extracted through lexical matching and analyzed across multiple textual hierarchically organized segments – words, sentences, paragraphs, chapters, literary works and the full corpus, which enabled the exploration of both local variation and global patterns of emotional signals. The approach combines emotional words density and intensity measures with hierarchical and chronological grouping, to examine short- and long-term changes in emotional structure across the corpus. First empirical results showed a stable baseline dominated by **trust** and **anticipation** emotions, alongside with the significantly increased emotional variation in the early twentieth century. These findings indicate that emotional organization in Serbian prose evolves in both magnitude and structural complexity over time and that proposed methodology enables consistent comparison and clear computational visibility across literary work segments and time periods.

Keywords: emotions, Plutchik model, lexicon, Serbian literary corpus

Leksičko istraživanje emocionalne dinamike u TEI romanima: studija slučaja na korpusu *SrpELTeC*

Ranka Stanković, Milena Šošić

Sažetak

Ovaj rad predstavlja hijerarhijsku analizu emocionalne dinamike u korpusu srpskih romana *SrpELTeC*. Studija kombinuje TEI-kodirane tekstove u korpusu sa *EmoLex.SR* leksikonom emocionalnih reči za srpski jezik, zasnovanim na Plutčikovom modelu osnovnih emocija. Emocionalni signali iz korpusa izdvojeni su putem leksičkog podudaranja po lemi i vrsti reči, a zatim su analizirani na više hijerarhijski organizovanih nivoa teksta, kao što su reči, rečenice, pasusi, poglavlja, književna dela i celokupni korpus. Ovako strukturisana analiza omogućila je istraživanje i prepoznavanje lokalnih varijacija i globalnih obrazaca emocionalnih signala u književnim tekstovima. Pristup kombinuje mere gustine emocionalnih reči i njihovog intenziteta sa hijerarhijskim i hronološkim grupisanjem, radi ispitivanja kratkoročnih i dugoročnih promena u emocionalnoj strukturi *SrpELTeC* korpusa. Prvi empirijski rezultati pokazali su stabilan osnovni obrazac u kome dominiraju emocije poverenje (eng. *trust*) i iščekivanje (eng. *anticipation*), uz značajno veći emocionalnu raznolikost u delima sa početka dvadesetog veka. Ovi nalazi ukazuju na to da se emocionalna organizacija u srpskoj prozi razvijala u pogledu intenziteta i strukturne složenosti tokom vremena, kao i da je predložena metodologija u ovom radu omogućila računarsku interpretabilnost emocionalnog aspekta jezika u različitim segmenatima pojedinačnih književnih dela, kao njihovo međusobno poređenje na osnovu vremenskih perioda u kojima su nastala.

Ključne reči: emocije, Plutčikov model, leksikon, srpski književni korpus

Comparative Intrinsic Evaluation of LLM Tokenizers on Serbian Text

Mihailo Škorić¹[0000-0003-4811-8692], Nikola Janković²[0000-0003-3484-4220],
Tanja Samardžić³[0000-0001-6451-3946]

¹ University of Belgrade, Faculty of Mining and Geology, Serbia,
mihailo.skoric@rgf.bg.ac.rs

² University of Belgrade, Faculty of Philology, Serbia,
nikolajankovickv@gmail.com

³ IDSIA USI-SUPSI, Lugano, Switzerland, tanja.samardzic@supsi.ch

Abstract

This study presents a comparative intrinsic evaluation of large language model tokenizers trained specifically for Serbian, a language characterized by rich morphology and dual script usage. We construct several tokenizers using diverse strategies, including GPT2-style and trailing space BPE, as well as some experimental approaches: minimum redundancy exploration, casing and script compression, and morphology-aware tokenization (based on a Serbian morphological dictionary). In addition, we benchmark tokenizers from contemporary models such as Qwen3.5 and Mistral3.5. Training is conducted on a Serbian corpus of approximately one billion words, with evaluation performed on a held out set one-third of that size. Intrinsic properties are analysed using Tokenizer Analysis Suite, calculating metrics such as fertility, compression ratio, Rényi Entropy and MorphScore. Our results highlight systematic differences different tokenizer building strategies: morphology driven approaches achieve higher MorphScore, while script compression strategies result in lower fertility and improved compression, especially while handling mixed Cyrillic–Latin corpora. In comparison, existing general purpose tokenizers reveal that they are consistently inferior to language specific tokenizers across most intrinsic metrics, underscoring the importance of tailored tokenizer design.

This work provides the initial comparative study of Serbian tokenizers, offering insights into how design choices affect the intrinsic properties. By systematically evaluating both newly trained and existing models, we establish a foundation for future research on tokenizer development for morphologically rich languages, and contribute benchmarks that can help future designing of tailored, specific purpose tokenizers.

Keywords: Tokenization, Serbian, Intrinsic evaluation, LLMs

Komparativna intrinzična evaluacija LLM tokenizatora na srpskom tekstu

Mihailo Škorić, Nikola Janković, Tanja Samardžić

Sažetak

Ovaj rad predstavlja uporednu intrinzičnu evaluaciju tokenizatora za velike jezičke modele, obučenih posebno za srpski jezik, koji karakteriše bogata morfologija i upotreba dva pisma. Konstruisali smo nekoliko tokenizatora koristeći različite strategije, uključujući standardno BPE u GPT2 stilu i BPE sa pratećim razmakom, kao i neke eksperimentalne pristupe: istraživanje minimalne redundantnosti, kompresiju pisam i kapitalizacije, i tokenizaciju potpomognutu morfologijom (na osnovu srpskog morfološkog rečnika). Pored toga, upoređujemo tokenizatore sa savremenim modelima kao što su Qwen3.5 i Mistral3.5. Obučavanje svih tokenizatora se sprovodi na srpskom korpusu od približno milijardu reči, a evaluacija se vrši na odvojenom, tri puta manjem skupu. Intrinzične osobine se analiziraju pomoću softvera Tokenizer Analysis Suite, izračunavajući metrike kao što su fertilitet, stepen kompresije, Renijeva entropija i morfoskor. Naši rezultati ističu sistematske razlike u rezultatima postignutim različitim strategijama izgradnje tokenizatora: pristupi vođeni morfologijom postižu veći morfskor, dok strategije kompresije postižu niži fertilitet i poboljšanu kompresiju, posebno pri radu sa mešovitim ćirilčno-latiničnim korpusima. U poređenju, postojeći tokenizatori opšte namene pokazuju da su konstantno po većini intrinzičnih metrika inferiorniji u odnosu na tokenizatore specijalizovane za jedan jezik, što naglašava važnost prilagođenog dizajna tokenizatora.

Ovaj rad pruža početnu uporednu studiju srpskih tokenizatora, nudeći uvid u to kako izbori strategija utiču na intrinzična svojstva. Sistematskim upoređivanjem postojećih i novoobučenih modela, postavljamo osnovu za buduća istraživanja razvoja tokenizatora za morfološki bogate jezike i doprinosimo benčmarkingu koji će pomoći u dizajniranju prilagođenih tokenizatora za specifične namene u budućnosti.

Ključne reči: Tokenizacija, Srpski, Intrinzična evaluacija, VJM

Wiki Dumps to Training Corpora: South Slavic Case

Mihailo Škorić¹[\[0000-0003-4811-8692\]](#), Cosimo Palma²[\[0000-0002-8161-9782\]](#)

¹ University of Belgrade, Faculty of Mining and Geology, Serbia,
mihailo.skoric@rgf.bg.ac.rs

² University of Pisa, Department of Computer Sciences, Italy,
cosimo.palma@phd.unipi.it

Abstract

This paper presents a pipeline designed to transform raw Wikimedia dumps into quality textual corpora for seven South Slavic languages. The work is divided into two major phases. The first involves extracting and cleaning usable natural language text from raw dumps of Wikipedia, Wikisource, Wikibooks, Wikinews, and Wikiquote. The second phase addresses the challenge of questionable or low-quality articles, which are often generated from databases or structured knowledge bases. These articles are characterised by repetitive patterns, generic phrasing, and minimal to no original content. To mitigate their impact, an n-gram-based filtering strategy is employed to detect high levels of textual redundancy between articles to be hence removed from the corpora entirely. The resulting datasets support language model training and comparative research across South Slavic languages, though the pipeline itself is largely language-agnostic and generalises beyond them.

Keywords: Text corpora, Wiki projects, Data cleaning, Data quality

Od Viki izvoza do korpusa za obučavanje: Slučaj južnoslovenskih jezika

Mihailo Škorić, Cosimo Palma

Sažetak

Ovaj rad predstavlja proces osmišljen da transformiše sirove Vikimedijine dampove u kvalitetne tekstualne korpusa za sedam južnoslovenskih jezika. Rad je podeljen u dve glavne faze. Prva uključuje ekstrakciju i čišćenje teksta iz sirovih izvoza Vikipedije, Vikizvornika, Vikiknjiga, Vikivesti i Wikicitata. Ovaj korak zahteva pažljivo rukovanje sirovim viki oznakama kako bi se izolovali, pre svega, tekstualni članci, a zatim i upotrebljiv tekst na prirodnom jeziku unutar njih.

Druga faza se bavi izazovom sumnjivih ili nekvalitetnih članaka, koji se često generišu iz baza podataka ili strukturiranih baza znanja. Ove članke karakterišu ponavljajući obrasci, generičko fraziranje i minimalan ili nikakav originalni sadržaj. Da bi se ublažio njihov uticaj, korišćena je strategija filtriranja zasnovana na n-gramima kako bi se otkrili visoki nivoi tekstualne redundantnosti među člancima, pri čemu se detektovani članci u potpunosti uklonjaju iz korpusa. Dobijeni skupovi podataka imaju za cilj da obezbede lingvistički bogate tekstove pogodne za obuku jezičkih modela ili sprovođenje komparativnih istraživanja na južnoslovenskim jezicima.

Kombinovanjem sistematske ekstrakcije sa kontrolom kvaliteta, ovaj rad doprinosi stvaranju pouzdanih, visokoinformativnih korpusa koji odražavaju autentične kulturne kontekste jezika. Iako se rad fokusira na slučaj južnoslovenskih jezika, pristup je uglavnom jezički nezavistan i može se primeniti na druge jezike i porodice jezika.

Ključne reči: Tekstualni korpusi, Viki projekti, Čišćenje podataka, Kvalitet podataka

Language Technologies

Поредбени фразеологизми са компонентом *нас* у српском језику: корпусно заснована анализа

Маријана Ђукић¹[0000-0003-3767-4247],
Слободан Новокмет¹[0000-0002-8525-5680]

¹ Српска академија наука и уметности, Институт за српски језик, Србија,
{marijana.bogdanovic|slobodan.novokmet}@isj.sanu.ac.rs

Abstract

У раду се анализирају поредбени фразеологизми српскога језика са зоокомпонентом *нас*, ексцерпирани из два корпуса српског језика: општег и прилично избалансираног *Корпуса савременог српског језика (СрпКор2021+)* и веб-корпуса *CLASSLA-WEB.SR 2.0*. На основу анализе корпусне грађе утврђен је инвентар поредбених фразеологизама са компонентом *нас* који се налазе у актуелној језичкој употреби, те сагледане њихове структурне и семантичке карактеристике, варијантност, као и модификације, чиме је добијен увид у начин реализације предметних фразеологизама у савременом српском језику. Добијена корпусна слика фразеологизама упоређена је са њиховим описом у речницима, општим и специјалним фразеолошким. Установљене су неподударности у начину на који су поједини фразеологизми лексикографски обрађени, пре свега у погледу лематизације и семантичког описа. Такође, издвојени су и фразеологизми који су у корпусу потврђени као устаљени, а које прегледани лексикографски извори не бележе. Добијени налази указују на чињеницу да постојећи лексикографски извори не осликавају у потпуности актуелну употребу фразеологизама, те истичу потребу за израдом фразеолошког речника заснованог на аутентичном корпусном материјалу.

Кључне речи: српски језик, фразеологија, лексикографија, електронски корпуси, поредбени фразеологизми

Comparative Phraseological Units with the Component *dog* in the Serbian Language: A Corpus-Based Analysis

Marijana Đukić, Slobodan Novokmet

Summary

This paper examines Serbian comparative phraseological units containing the zoonymic component *pas* ‘dog’. The data were extracted from two Serbian-language corpora: the Corpus of Contemporary Serbian (*Srp-Kor2021+*) and the *CLASSLA-web.sr 2.0* web corpus. The corpus analysis established an inventory of phraseological units attested in contemporary usage and examined their structural and semantic properties, variation, and modifications. The corpus findings were then compared with their treatment in general and specialised phraseological dictionaries. The analysis revealed discrepancies in variation, lemmatisation, and semantic description, as well as conventionalised expressions not recorded in the dictionaries consulted. The results show that existing lexicographic sources do not fully reflect current phraseological usage and point to the need for a corpus-based phraseological dictionary.

Key words: Serbian language, phraseology, lexicography, electronic corpora, comparative phraseological units.

Именице и колокације придев–именица у српском делу паралелног корпуса часописа *Инфотека*: текстометријски приступ

Јелена Андоновски¹[0009-0007-9764-5364],
Александра Тртовац¹[0000-0002-0478-9980]

¹ Универзитетска библиотека “Светозар Марковић”, Србија,
{andonovski|aleksandra}@unilib.rs

Апстракт

Овај рад приказује фреквенцијску анализу именица и придевско-именичких колокација, као и текстометријску анализу (анализа прогресије и специфичности термина) српског дела доменског паралелног корпуса научног часописа *Инфотека*, *Infoteka-sr*. Фокус истраживања је на поређењу два периода: фазе усмерене на библиотекарство и информатику (до 2014) – поткорпус VI и фазе окренуте дигиталној хуманистици (од 2016) – поткорпус DH. За текстометријску анализу коришћен је алат TХМ, а добијени резултати указују на промену тематског фокуса часописа, од традиционалних библиотечко-информационих појмова ка концептима дигиталне хуманистике, језичких ресурса и технологија. Корпус *Infoteka-sr* потврђује се као релевантан дигитални ресурс за даља истраживања у области библиотекарства и информатике и дигиталне хуманистике.

Кључне речи: Инфотека, текстометрија, библиотекарство и информатика, дигитална хуманистика, корпус, језички ресурси, фреквенцијска анализа

Nouns and Adjective–Noun Collocations in the Serbian Subcorpus of the Infoteka Parallel Corpus: A Textometric Analysis

Jelena Andonovski, Aleksandra Trtovac

Abstract

This paper presents textometric analyzes Serbian part of the domain-specific parallel corpus of the scientific journal Infotheca, Infoteka-sr. The research focuses on comparing two periods: the phase oriented toward library and information science (up to 2014) – subcorpus BI and the phase focused on digital humanities (from 2016 onward) – subcorpus DH. Using the TXM tool, a frequency analysis of nouns and adjective–noun collocations was conducted, along with a specificity analysis of terms. The results indicate a shift in the journal’s thematic focus, from traditional library and information science concepts toward concepts related to digital humanities, language resources, and technologies. The Infoteka-sr corpus is confirmed as a relevant digital resource for further research in the fields of library and information science and digital humanities.

Keywords: Infotheca, textometry, library and information science, digital humanities, corpus, language resources, frequency analysis.

Hypothesis-Driven Corpus Querying at Scale: A Method and Its Application to *-nje* Nominalizations in BCMS

Predrag Kovačević¹[\[0000-0003-2545-9240\]](#), Marko Simonović²[\[0000-0002-9651-6399\]](#), Marija Đokić Petrović³[\[0000-0002-0568-5219\]](#)

¹ University of Novi Sad, Faculty of Philosophy, Department of English Studies, Serbia, pedjakovacevic90@gmail.com;

² University of Graz, Department of Slavic Studies, Austria, marko.simonovic@uni-graz.at;

³ Union University, School of Computing, Belgrade, Serbia, mdjokicpetrovic@raf.rs

Abstract

This paper presents a reproducible method for hypothesis-driven corpus querying at scale and applies it to Bosnian/Croatian/Montenegrin/Serbian (BCMS) *-nje* nominalizations. Using pluralization and genitive complements as diagnostics of internal verbal structure, the study tests predictions concerning structural differences between perfective and imperfective deverbal nominals across four BCMS varieties. The paper introduces a fully automated workflow for large-scale corpus querying, enabling systematic testing of theoretically motivated grammatical hypotheses across 24,000 corpus queries. The results reveal stable asymmetries between perfective and imperfective nominalizations consistent with previous theoretical proposals, while also uncovering limited but systematic microvariation across BCMS varieties. Beyond its empirical contribution, the study provides a generalizable methodological framework for scalable and reproducible corpus-based grammatical research.

Keywords: nominalization, Slavic linguistics, BCMS, aspect, corpus linguistics, reproducibility

Korpusno pretraživanje velikog obima zasnovano na hipotezi: Metod i njegova primena na glagolske imenice na *-nje* u bchs-u

Predrag Kovačević, Marko Simonović, Marija Đokić Petrović

Sažetak

Rad predstavlja reproducibilan metod za korpusno pretraživanje velikog obima zasnovano na hipotezama te njegovu primenu na nominalizacije na *-nje* u bosanskom/crnogorskom/hrvatskom/srpskom jeziku. Polazeći od pluralizacije i genitivnih dopuna kao dijagnostika unutrašnje glagolske strukture, rad ispituje razlike između perfektivnih i imperfektivnih glagolskih imenica. Rezultati pokazuju stabilne asimetrije: imperfektivne nominalizacije ređe se pluralizuju i češće se javljaju sa genitivnim dopunama, dok perfektivne nominalizacije pokazuju suprotnu tendenciju. Time se potvrđuju ranije teorijske pretpostavke i pokazuje da automatizovano korpusno pretraživanje omogućava pouzdano i reproducibilno testiranje gramatičkih hipoteza.

Ključne reči: nominalizacija, slovenska lingvistika, bchs, glagolski vid, korpusna lingvistika, reproducibilnost

СентИроСа: корпус за детекцију ироније и сарказма у кратким текстовима на српском језику

Данка Јокић¹[0000-0002-7621-5847] , Милица Андрић²[0009-0008-0319-4548]

¹ Универзитет у Београду, Србија, danka.jokic@rcub.bg.ac.rs

² Универзитет у Београду, Филолошки факултет, Србија, milica@mail.ru

Апстракт

Препознавање ироније и сарказма у тексту може побољшати перформансе многих система за обраду природног језика (ОПЈ), укључујући сумирање текста, анализу сентимента, откривање увредљивог језика итд. Развој система за аутоматску детекцију ироније и сарказма захтева ручно обележене скупове података за обучавање и евалуацију модела. Међутим, за српски језик до сада није постојао јавно доступан ручно обележен корпус потребан за аутоматску детекцију ироније и сарказма. У овом раду представљамо СентИроСа, први јавно доступан ручно аотирани корпус за детекцију ироније и сарказма на српском језику. Корпус је формиран проширивањем постојеће колекције кратких текстова са платформе Твитер/Х, претходно аотираних ознакама сентимента, новим ручним аотацијама за иронију и сарказам. Састоји се од 1117 твитова која су обележавала три независна аотатора. У раду су описани поступак аотације и резултати квалитативне и квантитативне анализе корпуса, укључујући анализу слагања аотатора, лингвистичких маркера ироније и сарказма и повезаности ових феномена са сентиментом. СентИроСа представља нови језички ресурс за истраживања ироније, сарказма и других повезаних феномена у обради природног језика на српском језику.

Кључне речи: иронија, сарказам, корпуси

SentIroSa: A Corpus for Irony and Sarcasm Detection in Short Serbian Texts

Danka Jokić, Milica Andrić

Abstract

The detection of irony and sarcasm in text can improve the performance of many Natural Language Processing (NLP) systems, including text summarization, sentiment analysis, and abusive language detection. The development of automatic irony and sarcasm detection systems requires manually annotated datasets for model training and evaluation. However, until now, no publicly available manually annotated corpus for automatic irony and sarcasm detection in Serbian has been available. In this paper, we present SentIroSa, the first publicly available manually annotated corpus for irony and sarcasm detection in Serbian. The corpus was created by extending an existing collection of short texts from Twitter/X, previously annotated for sentiment, with new manual annotations for irony and sarcasm. It consists of 1,117 tweets annotated by three independent annotators. We describe the annotation procedure and present the results of qualitative and quantitative analyses of the corpus, including inter- and intra-annotator agreement, linguistic markers of irony and sarcasm, and the relationship between these phenomena and sentiment. SentIroSa provides a new language resource for research on irony, sarcasm, and other related phenomena in Serbian NLP.

Key words: irony, sarcasm, corpus

ADMIRE 2 Serbian: Изградња српског мултимодалног скупа идиома

Данка Јокић^{1,*}[0000-0002-7621-5847], Срђан Шућур^{2,*}[0009-0003-6815-2887],
Јелена Марковић²[0000-0003-4962-3754]

¹ Универзитет у Београду, Србија, danka.jokic@rcub.bg.ac.rs

² Универзитет у Источно Сарајеву, Филозофски факултет Пале, Катедра за англистику,
{[srdjan.sucur](mailto:srdjan.sucur@ffuis.edu.ba)|[jelena.markovic](mailto:jelena.markovic@ffuis.edu.ba)}@ffuis.edu.ba

Апстракт

У овом раду представљен је српски мултимодални скуп идиома и процес његове изградње, који је реализован у оквиру *Admire 2* иницијативе. Размотрени су изазови припреме српских еквивалената за енглеске вишечлане идиоматске изразе и примера њихове употребе, као и њихове визуелне репрезентације помоћу генеративног модела за синтезу слика *Midjourney*. Кад год је то било могуће, примери употребе прикупљани су из језичких корпуса и других текстуалних извора како би се обезбедили аутентичност, природност и језичка валидност представљених израза. Посебна пажња посвећена је проналажењу одговарајућих превода идиома усклађених са српским културним наслеђем. Приликом генерисања слика уочена су ограничења генеративног модела у приказивању апстрактних појмова, сложених емоционалних стања и просторних односа, уз присуство различитих облика пристрасности у генерисаним приказима.

Кључне речи: идиоми, мултимодални корпуси, пристрасност АИ модела

ADMIRE 2 Serbian: Building a Serbian Multimodal Idiom Dataset

Danka Jokić, Srđan Šućur, Jelena Marković

Summary

This paper presents a Serbian multimodal idiom dataset and the process of its construction, developed within the *Admire 2* initiative. It discusses the challenges involved in preparing Serbian equivalents of English multi-word idiomatic expressions and examples of their use, as well as their visual representations generated using the *Midjourney* generative image synthesis model. Whenever possible, usage examples were collected from language corpora and other textual sources to ensure the authenticity, naturalness, and linguistic validity of the presented expressions. Particular attention was devoted to identifying appropriate idiom translations aligned with Serbian cultural heritage. The image generation process revealed limitations of the generative model in depicting abstract concepts, complex emotional states, and spatial relationships, as well as various forms of bias in the generated images.

Keywords: idiomatic expressions, multimodal corpora, model bias

AI4LANG – Integrating Artificial Intelligence for the Digital Transformation of Language Teaching

Benedikt Perak¹[\[0000-0003-4177-5307\]](#), Dragana Špica²[\[0000-0003-1008-2700\]](#)

¹ University of Rijeka, Faculty of Humanities and Social Sciences / Syntagent d.o.o., Rijeka, Croatia, bperak@uniri.hr

² University of Jurja Dobrile in Pula, Faculty of Humanities, Croatia, dragana.spica@unipu.hr

Abstract

The rapid advancement of generative AI and natural language processing has created transformative opportunities for language education, yet the integration of these technologies into pedagogically coherent, CEFR-aligned learning environments remains a significant challenge. This paper presents AI4LANG, a multinational research and development initiative that expands a production-ready AI Language Tutor platform across three typologically distinct languages – English, Japanese, and Croatian – addressing the global need for adaptive, AI-mediated language instruction that serves both high-resource and underrepresented languages. The platform’s architecture is built on a multi-agent orchestration paradigm where specialized AI agents handle distinct pedagogical functions – exercise generation, pronunciation evaluation, grammar feedback, and learning path adaptation – coordinated through a central orchestrator operating over a dual-database knowledge layer combining vector embeddings with structured lexical and grammatical graphs. We present the technical architecture, the CEFR-aligned pedagogical framework underpinning adaptive exercise generation, and the methodology for creating language-specific knowledge resources through LLM-assisted inference validated by expert review. The paper contributes a replicable model for developing multilingual AI tutoring systems that balance technological sophistication with pedagogical rigor across diverse linguistic contexts.

Keywords: AI language tutoring, multi-agent architecture, CEFR, multilingual NLP, adaptive learning, knowledge graphs, speech recognition

AI4LANG – Integracija umjetne inteligencije za digitalnu transformaciju nastave jezika

Benedikt Perak, Dragana Špica

Sažetak

Brz napredak generativne umjetne inteligencije i obrade prirodnog jezika otvorio je mogućnosti za korjenito unaprjeđenje nastave jezika, no integracija tih tehnologija u pedagoški usklađena okruženja za učenje, utemeljena na Zajedničkom europskom referentnom okviru za jezike (CEFR), i dalje predstavlja značajan izazov. U ovom se radu predstavlja AI4LANG, međunarodna istraživačko-razvojna inicijativa koja proširuje platformu AI Language Tutor, spremnu za produkcijsku uporabu, na tri tipološki različita jezika – engleski, japanski i hrvatski – odgovarajući na globalnu potrebu za prilagodljivom nastavom jezika posredovanom umjetnom inteligencijom, namijenjenom i jezicima s bogatim resursima i nedovoljno zastupljenim jezicima.

Arhitektura platforme temelji se na paradigmi višeagentne orkestracije, u kojoj specijalizirani agenti umjetne inteligencije obavljaju različite pedagoške funkcije – generiranje vježbi, procjenu izgovora, pružanje povratnih informacija o gramatici i prilagodbu putanje učenja. Njihov rad koordinira središnji orkestrator koji djeluje nad slojem znanja utemeljenim na dvjema bazama podataka, kombinirajući vektorske reprezentacije sa strukturiranim leksičkim i gramatičkim grafovima.

Predstavljamo tehničku arhitekturu, pedagoški okvir usklađen sa CEFR-om na kojem se temelji prilagodljivo generiranje vježbi te metodologiju izrade jezično specifičnih resursa znanja zaključivanjem uz pomoć velikih jezičnih modela (LLM), čiji se rezultati potvrđuju stručnom provjerom. Doprinos rada jest model za razvoj višejezičnih sustava za poučavanje utemeljenih na umjetnoj inteligenciji koji se može primijeniti i u drugim sredinama, a koji u različitim jezičnim kontekstima uravnotežuje tehnološku naprednost i pedagošku utemeljenost.

Ključne riječi: Poučavanje jezika uz pomoć umjetne inteligencije, višeagentna arhitektura, CEFR, višejezična obrada prirodnog jezika, prilagodljivo učenje, grafovi znanja, prepoznavanje govora

The ItalSerb Spoken Corpus of L2 Italian and a Paired Subcorpus of L1 Serbian

Nevena Ceković¹[0000-0002-5407-5759]

¹University of Belgrade, Faculty of Philology, Department of Italian Studies,
Serbia, n.cekovic@fil.bg.ac.rs

Abstract

This paper presents **progress in the development of** a subcorpus of spoken production in the native language (L1) of Serbian-speaking students of Italian as a foreign language (L2), paired with the main ItalSerb corpus of L2 Italian. The subcorpus consists of 46 samples of dialogic and monologic production, amounting to approximately four hours of recordings, collected on two occasions from 23 students at an intermediate level of L2 proficiency. Particular attention is devoted to the methodological considerations underlying the construction of the subcorpus, including participant data, interaction type, genre and topic, with special emphasis on matching the newly collected metadata with the existing metadata included in the ItalSerb collection. The preliminary analysis of the L1 and L2 production of one participant followed longitudinally reveals similarities and differences in the patterns and frequency of use of the pragmatic elements under investigation, namely discourse markers functioning in **turn initiation (floor-taking)**. Idiosyncratic preferences for particular marker forms that are not functionally equivalent across the two languages were also identified. The significance of these findings lies in their potential contribution to research on the mutual influence of L2 and L1. The methodological experience gained in constructing the subcorpus also provides a basis for developing guidelines for the digitisation and transcription of the collected (sub)corpus data.

Keywords: ItalSerb, learner corpus, paired corpus, L2 Italian, L1 Serbian, Serbian-speaking students, discourse markers

ItalSerb: govorni korpus L2 italijanskog i upareni subkorpus L1 srpskog

Nevena Ceković

Sažetak

U radu se predstavljaju dostignuća u izgradnji uparenog subkorpusa govorne produkcije na maternjem jeziku (L1) srbofonih studenata italijanskog kao stranog (L2). Glavni korpus govorne produkcije na L2 italijanskom, ItalSerb, time je dopunjen uparenim subkorpusom produkcije na L1 srpskom, prikupljene od istih ispitanika tokom odgovarajućeg perioda studiranja. Učenički korpus ItalSerb (ital. *ITALiano dei SRBofoni* – ITALijanski SRBofonih), prema našim saznanjima, prvi je i jedini korpus L2 italijanskog u Srbiji, gde do sada nije zabeleženo ni postojanje uparenih učeničkih zbirki za italijanski kao strani jezik. Obuhvata 28 sati govornog materijala i 200.000 reči u ortografskim transkriptima, odnosno 315 audio-uzoraka produkcije 172 srbofona studenta italijanskog, praćenih transverzalno i longitudinalno tokom četiri studijske godine i na različitim ciljnim nivoima jezičke kompetencije (od A2+ do C1). Relevantnost kolekcije naročito dolazi do izražaja s obzirom na mali broj učeničkih korpusa za druge strane jezike, kao i na mali broj drugih italijanističkih korpusnih zbirki, nastalih u Srbiji. Značajna je i u međunarodnom kontekstu korpusa specijalizovanih za italijanski kao strani (među kojima je, prema dostupnim podacima, po ukupnom trajanju snimljenog materijala veći jedino švedski InterIta, sa približno 30 sati snimaka) ili kao drugi jezik (LIPS, Banca dati di italiano L2, VIP), kreiranih u italijanskim i drugim inostranim univerzitetskim centrima.

Novoformirani subkorpus produkcije na L1 srpskom čini 46 uzoraka dijaloške i monološke produkcije, ukupnog trajanja približno četiri sata (prosečno oko pet minuta po uzorku). Uzorci su prikupljeni u dva navrata, u razmaku od šest meseci, od 23 studenta na srednjem nivou kompetencije u L2, tokom petog i šestog semestra studija, kada su ciljni nivoi kompetencije u L2 bili u rasponu od B1 do B2+. Interakcija je elicirana, kao i u korpusu ItalSerb, po modelu certifikacionih testova, na raznovrsne teme i u vidu asimetričnog polustrukturisanog intervjua. Ispitanici su za prvi zadatak uglavnom birali omiljenu muziku ili hranu, a za drugi opis jednog letovanja. Svaki ispitanik je u oba L1 snimanja obrađivao iste teme. Uporedivost teme, tipa zadatka i žanra produkcije u L1 i L2 omogućavaju identifikovanje karakterističnih jezičkih obrazaca i sistematsko poređenje upotrebe u maternjem i stranom jeziku.

Posebna pažnja u radu posvećena je metodološkim pretpostavkama u izgradnji subkorpusa: pseudonimizaciji i zaštiti podataka, tipu, žanru i tematici interakcije, kao i uparivanju novoprikupljenih metapodataka iz subkorpusa sa postojećim podacima kolekcije ItalSerb. Preliminarna analiza produkcije jednog longitudinalno praćenog ispitanika na L1 i L2, sprovedena u radu, ukazuje na sličnosti i razlike u obrascima i broju pojavljivanja diskursnih markera u funkciji uzimanja reči (pa u L1 i funkcionalno ekvivalentan združeni oblik *eh sì* u L2). Uočene su i idiosinkratičke preferencije prema pojedinim oblicima markera koji u dva jezika nisu funkcionalno ekvivalentni. Analiza odabrane markerske funkcije pruža početni uvid u odnos između produkcije na L1 i L2 i otvara prostor za detaljnije proučavanje međujezičkog uticaja, kako u vezi sa markerima diskursa tako i sa drugim jezičkim fenomenima.

Pored pitanja vezanih za koncipiranje subkorpusa L1 i njegovo povezivanje sa glavnim korpusom L2, uključujući zaštitu ličnih podataka, obradu metapodataka i sociolingvističkih informacija, u daljem radu potrebno je razmotriti i postupke ortografske transkripcije i digitalizacije prikupljene građe. Dosadašnja iskustva u tom pogledu mogu poslužiti kao osnova za izradu smernica za njenu dalju obradu i korišćenje. Predstavljanje osnovnih metodoloških principa izgradnje subkorpusa i povezivanja zbirki na L1 i L2, uz njihovu ilustraciju reprezentativnim korpusnim primerima, ima za cilj da ukaže na metodološki relevantne postupke u izradi učeničkih (sub)korpusa, kao i da ponudi smernice za njihovu primenu i korišćenje u različitim vrstama istraživanja u oblasti usvajanja L2 i didaktike stranih jezika.

Ključne reči: ItalSerb, učenički korpus, upareni korpus, L2 italijanski, L1 srpski, srbofoni studenti, diskursni markeri

Linguistic Annotation of South Slavic Languages with Large Language Models

Katerina Zdravkova¹[\[0000-0002-9674-3081\]](#)

¹ Ss. Cyril and Methodius University in Skopje, Faculty of Computer Science and Engineering, North Macedonia, katerina.zdravkova@finki.ukim.mk

Abstract

This study investigates the applicability of ChatGPT 5 Plus to linguistic annotation in four South Slavic languages: Bulgarian, Macedonian, Serbian, and Slovene. While recent studies have demonstrated the potential of large language models (LLMs) for linguistic annotation, their ability to learn detailed language-specific morphosyntactic annotation conventions remains insufficiently understood. Using the corresponding translations of the first 1235 tokens of Orwell’s 1984 together with their MULTEXT-East annotations, we evaluated the model for lemmatization, part-of-speech (PoS) tagging, and morphosyntactic description (MSD) annotation. The evaluation consisted of zero-shot annotation, one-shot calibration, iterative calibration through corrective human feedback, and an independent blind evaluation. Zero-shot annotation demonstrated high lemmatization accuracy and reliable PoS tagging, while iterative calibration substantially improved adherence to language-specific annotation conventions, particularly for MSD annotation. Blind evaluation on previously unseen sentences confirmed high lemmatization (94.95–99.22%) and PoS-tagging accuracy (94.69–96.61%) across all four languages, with MSD accuracy ranging from 75.16% to 89.21%. Most observed errors reflected the model’s tendency to rely on own linguistic knowledge rather than consistently adopting the language-specific annotation conventions introduced during calibration. Regardless of their relatively frequent presence, especially during MSD annotation, the current results show that this chatbot can effectively support morphosyntactic annotation using iterative calibration based on the human-in-the-loop concept and highlight the importance of expert feedback for adapting large language models to corpus-specific annotation standards.

Keywords: lemmatization, LLMs, MSD annotation, PoS tagging

Lingvistička anotacija južnoslovenskih jezika pomoću velikih jezičkih modela

Katerina Zdravkova

Sažetak

Ova studija istražuje primenljivost ChatGPT 5 Plus za lingvističku anotaciju četiri južnoslovenska jezika: bugarskog, makedonskog, srpskog i slovenačkog. Iako su nedavna istraživanja pokazala potencijal velikih jezičkih modela za lingvističku anotaciju, njihova sposobnost da usvoje detaljne jezički specifične konvencije morfosintaksičke anotacije još uvek nije dovoljno istražena. Koristeći odgovarajuće prevode prvih 1235 tokena iz Orvelovog romana „1984“, zajedno sa njihovim MULTEXT-East anotacijama, ocenili smo model za lematizaciju, označavanje vrsta reči (PoS) i anotaciju morfosintaksičkog opisa (MSD). Eksperiment je obuhvatio četiri faze: anotaciju bez ijednog pokušaja, kalibraciju sa jednokratnim učenjem zasnovanom na ručno anotiranim primerima, iterativnu kalibraciju putem korektivnih povratnih informacija i nezavisnu slepu evaluaciju. Anotacija bez ijednog pokušaja je već pokazala visoku tačnost lematizacije i pouzdano PoS označavanje, dok je iterativna kalibracija značajno poboljšala usklađenost sa jezički specifičnim konvencijama anotacije, naročito za MSD anotaciju. Slepa evaluacija na prethodno neviđenom skupu podataka potvrdila je visoku tačnost lematizacije (94.95–99.22%) i PoS označavanja (94.69–96.61%) u sva četiri jezika, dok se tačnost MSD anotacije kretala od 75.16% do 89.21%. Većina primećenih grešaka odražavala je tendenciju modela da se oslanja na sopstveno lingvističko znanje, umesto da dosledno usvaja jezičke specifične konvencije anotacije uvedene tokom kalibracije. Bez obzira na njihovu relativno čestu pojavu, naročito tokom MSD anotacije, dobijeni rezultati pokazuju da ovaj četbot može efikasno podržati morfosintaksičku anotaciju primenom iterativne kalibracije zasnovane na konceptu čovek u petlji i ističu značaj stručnih povratnih informacija za prilagođavanje velikih jezičkih modela korpusno specifičnim standardima anotacije.

Glavne reči: lematizacija, veliki jezički modeli, morfosintaksička anotacija, označavanje vrste reči

Anotacija moralnih vrednosti u srpskom onlajn diskursu: izazovi i perspektive

Jelena Redli¹[\[0000-0003-2518-5286\]](#)

¹Univerzitet u Novom Sadu, Filozofski fakultet,
Odsek za srpski jezik i lingvistiku, Srbija,
redli@ff.uns.ac.rs

Apstrakt

Rad se bavi problemima anotacije moralnih vrednosti u srpskom onlajn diskursu na primeru dela korpusa Corpus-Mor.SR-13, koji obuhvata objave sa Tvitera i Redita i sadrži oko 17.000 anotacija. U fokusu rada nisu samo anotacione kategorije i tehnički aspekti označavanja, već pre svega lingvistički i interpretativni problemi koji se javljaju u procesu anotacije kratkih, kontekstualno zavisnih i često implicitnih iskaza.

Polazi se od pretpostavke da moralna evaluacija u srpskom jeziku često nije izražena isključivo eksplicitnim leksičkim markerima, već da u njenom oblikovanju važnu ulogu mogu imati pragmatičke implikacije, ironija, izbor sintaksičkih struktura, diskursno pozicioniranje i kulturno prepoznatljivi obrasci vrednovanja. Posebna pažnja biće posvećena primerima u kojima dolazi do neslaganja među anotatorima, jer upravo takvi slučajevi mogu ukazati na granice površinske klasifikacije i na značaj šireg diskursnog i sociokulturnog konteksta. U radu će se razmatrati i specifičnosti pojedinih domena onlajn komunikacije, poput sportskog i političkog diskursa, u kojima se moralne vrednosti često iskazuju posredno, metaforički ili kroz kolektivne identitetske obrasce. Na osnovu analiziranog materijala nastojaće se da se ukaže na potrebu za razvojem detaljnijih smernica za anotaciju moralnih vrednosti u srpskom jeziku, kao i na značaj saradnje između lingvistike i savremenih pristupa obradi prirodnog jezika.

Ključne reči: anotacija korpusa, moralne vrednosti, onlajn diskurs, sportski diskurs, pragmatika, obrada prirodnog jezika, srpski jezik.

Moral Value Annotation in Serbian Online Discourse: Challenges and Perspectives

Jelena Redli

Summary

This paper examines linguistic and interpretative challenges in the annotation of moral values in Serbian online discourse. The analysis is based on a portion of the Corpus-Mor.SR-13 dataset, which contains annotated units from Twitter and Reddit, while the sports discourse subset is treated as a case study. The aim is not to provide a quantitative account of moral categories, but to show why moral value annotation cannot be reduced to explicit lexical markers or predefined keyword lists.

The theoretical framework combines Moral Foundations Theory with lexical-semantic, pragmatic and discourse-oriented approaches to meaning. The paper distinguishes moral evaluation from related forms of value-oriented analysis, especially sentiment analysis, emotion detection and the analysis of offensive or toxic online speech. While sentiment usually refers to the positive, negative or neutral orientation of an utterance, moral evaluation involves a more complex relation to actions, actors, social norms and values. This distinction is particularly important in short online texts, where context is often limited and evaluative meanings may remain implicit.

The sports discourse subset is analysed as a borderline domain in which neutral information, affective reaction and moral evaluation frequently overlap. The analysis focuses on several types of annotation problems: the boundary between the label *non-moral* and moral evaluation, the difference between sentiment and moral value, implicit moral evaluation, overlap between moral categories, and annotator disagreement.

The examples show that negative events such as defeat, injury or elimination may trigger an affective response without necessarily activating a moral category. Conversely, moral relevance may be established indirectly, through references to rules, institutional authority, suspected irregularity, injustice or collective loyalty. Annotator disagreement is therefore not treated only as a weakness of the annotation procedure, but also as linguistically relevant evidence of semantic and pragmatic instability.

The results indicate that Serbian annotation guidelines should more clearly distinguish unfavourable events, sentiment and moral evaluation. Although limited to sports discourse, the analysis is relevant for broader Serbian moral value annotation and for more reliable annotated resources.

Keywords: corpus annotation, moral values, online discourse, sports discourse, pragmatics, natural language processing, Serbian language

From Literary Memory to Platformed Culture: South Slavic Diaspora Identity and Algorithmic Marginalization in Canadian Multicultural Literature

Milica Jevtić¹[0009-0007-3188-9702], Tamara Vučenović²[0000-0003-4152-4372],
Miloš Stojanović¹[0009-0009-3774-6483], Danijela Nešić¹[0000-0001-9704-4701]

¹ University of Niš, Faculty of Applied Sciences, Serbia,
milengelma@gmail.com, milos.s87@hotmail.com, danijela.nesic.z@gmail.com

² Metropolitan University, Belgrade, Serbia, tamara.vucenovic@metropolitan.ac.rs

Abstract

This paper examines Canadian multicultural literature in relation to diaspora marginalization, focusing on Joy Kogawa's exploration of silence and erasure in *Obasan* (1981), Dionne Brand's fragmented constructions of identity in *A Map to the Door of No Return* (2001), and M. G. Vas-sanji's articulation of hybrid identity in *The In-Between World of Vikram Lall* (2003). These literary perspectives are then extended to the digital age through an analysis of South Slavic diaspora communities in Canada (Serbian, Croatian, and Bosnian) and their forms of online cultural production and engagement. In the digital era, cultural marginalization is no longer solely produced through colonial discourse, patriarchy, or state violence, but is increasingly intensified by algorithmic regimes of visibility, self-representation, and surveillance. The paper demonstrates that digital platforms have transformed diaspora identity from print-based cultural memory into a participatory and continuously evolving system shaped by algorithmic logics. Therefore, marginalization occurs less through direct exclusion from literary production and more through language loss, reliance on dominant digital platforms, and the replacement of traditional publishing with informal online storytelling. The aim of this paper is to demonstrate that digitization has not resulted in the democratization of the literary and linguistic sphere, but rather that algorithmic filtering has contributed to the marginalization of languages with lower economic capital, including South Slavic languages.

Keywords: Canadian multicultural literature, cultural marginalization, digital era, minority voices, diaspora identity, hybrid identity, algorithm.

Od književnog pamćenja do platformске kulture: identitet južnoslovenske dijaspore i algoritamska marginalizacija u kanadskoj multikulturalnoj književnosti

Milica Jevtić, Tamara Vučenović, Miloš Stojanović

Sažetak

Ovaj rad analizira kanadsku multikulturalnu književnost u odnosu na marginalizaciju dijaspore, fokusirajući se na istraživanje Džoj Kogave o tišini i potiskivanju u delu „Obasan“ (1981), fragmentarne konstrukcije identiteta u delu Dion Brand „Mapa do vrata bez povratka“ (2001) i artikulaciju hibridnog identiteta M. G. Vasandija u delu „Vikram Lal između dva sveta“ (2003). Ove književne perspektive se zatim proširuju na digitalno doba kroz analizu zajednica južnoslovenske dijaspore u Kanadi (srpske, hrvatske i bosanske) i njihovih oblika onlajn kulturne produkcije i angažovanja. U digitalnoj eri, kulturna marginalizacija više se ne proizvodi isključivo kroz kolonijalni diskurs, patrijarhat ili državno nasilje, već je sve više intenzivirana algoritamskim režimima vidljivosti, samoreprezentacije i nadzora. Rad pokazuje da su digitalne platforme transformisale identitet dijaspore iz kulturnog pamćenja zasnovanog na štampi u participativni i kontinuirano razvijajući sistem oblikovan algoritamskom logikom. Stoga se marginalizacija manje dešava kroz direktno isključivanje iz književne produkcije, a više kroz gubitak jezika, oslanjanje na dominantne digitalne platforme i zamenu tradicionalnog izdavaštva neformalnim onlajn pripovedanjem. Cilj ovog rada je da pokaže da digitalizacija nije dovela do demokratizacije književne i jezičke sfere, već da je algoritamsko filtriranje doprinelo marginalizaciji jezika sa nižim ekonomskim kapitalom, uključujući južnoslovenske jezike.

Ključne reči: Kanadska multikulturalna književnost, kulturna marginalizacija, digitalna era, glasovi manjina, identitet dijaspore, hibridni identitet, algoritmi.

Deverbal Nominalisations in the Timok Speech: A Preliminary Corpus Study

Andela Babić¹[\[0000-0002-7472-8198\]](#), Marko Simonović²[\[0000-0002-9651-6399\]](#),
Predrag Kovačević³[\[0000-0003-2545-9240\]](#)

¹ Serbian Academy of Sciences and Arts, Institute for Balkan Studies, Serbia,
andjela.redzic@bi.sanu.ac.rs

² University of Graz, Department of Slavic Studies, Austria,
marko.simonovic@uni-graz.at

³ University of Novi Sad, Faculty of Philosophy, Department of English Studies,
Serbia, predrag.kovacevic@ff.uns.ac.rs

Abstract

This paper presents a preliminary corpus-based study of deverbal nominalisations in the Timok speech, a speech of the Prizren-Timok dialect of Serbian. The analysis focuses on two productive morphological classes: nominalisations in *-nje/-će* and nominalisations in *-cija*. The study is based on 606 tokens extracted from the Spoken Torlak Dialect Corpus 1.0 and examines four properties: the aspectual class of the source verb, complement realisation, pluralisation, and correspondence with standard Serbian forms. The results show a strong predominance of *-nje/-će* nominalisations, especially those derived from imperfective verbs, while *-cija* nominalisations are considerably less frequent. Overt complements are rare overall, suggesting limited preservation of verbal argument structure in the corpus data, although they occur more often with *-nje/-će* nominalisations than with *-cija* forms. By contrast, plural forms are proportionally more frequent among *-cija* nominalisations, supporting the view that this class is more strongly lexicalised and more fully integrated into the nominal domain. Finally, most tokens show morphological correspondence with standard Serbian, although the data also leave room for further investigation of dialect-specific derivational patterns. The study contributes to research on deverbal nominalisation in South Slavic by extending the empirical basis to a non-standard Balkan Slavic variety and to spontaneous spoken-language data.

Keywords: deverbal nominalisations, Timok speech, Prizren-Timok dialect, corpus study, aspect, pluralisation, complements

Glagolske imenice u timočkom govoru: preliminarna korpusna studija

Anđela Babić, Marko Simonović, Predrag Kovačević

Sažetak

U radu se predstavlja preliminarna korpusna analiza glagolskih imenica u timočkom govoru, jednom od govora prizrenko-timočkog dijalekta srpskog jezika. Analiza je zasnovana na 606 primera dobijenih iz korpusa *Spoken Torlak Dialect Corpus 1.0* i fokusira se na dve produktivne morfološke klase: imenice na *-nje/-će* i imenice na *-cija*. Ispituju se četiri svojstva: vid glagola od kojeg je nominalizacija izvedena, realizacija dopuna, upotreba množine i morfološko poklapanje sa standardnim srpskim jezikom.

Rezultati pokazuju izrazitu dominaciju imenica na *-nje/-će*, naročito onih izvedenih od nesvršenih glagola. Imenice na *-cija* znatno su ređe u analiziranom materijalu. Izražene dopune pojavljuju se veoma retko, što upućuje na ograničeno očuvanje glagolske argumentske strukture u spontanoj govornoj građi. Ipak, kada su dopune potvrđene, one se uglavnom javljaju uz imenice na *-nje/-će*. Nasuprot tome, oblici množine proporcionalno su najčešći kod imenica na *-cija*, što potvrđuje pretpostavku da je ova klasa leksikalizovanja, brojiva i čvršće integrisana u imenički domen.

Većina potvrđenih oblika morfološki se poklapa sa odgovarajućim oblicima u standardnom srpskom jeziku. Ipak, raspoloživa građa ne omogućava pouzdane zaključke o tome da li je reč o uticaju standardnog jezika, zajedničkom nasleđenom obrascu ili unutar-dijalekatskoj tvorbi. Rad time proširuje istraživanje glagolskih imenica na nestandardni balkanoslovenski govor i na spontani govorni korpus, otvarajući prostor za dalja poređenja sa drugim prizrenko-timočkim govorima i sa govornim standardnim srpskim jezikom.

Cljučne reči: glagolske imenice, timočki govor, prizrenko-timočki dijalekat, korpusna studija, vid, množina, dopune

Exploring In-Context Learning for LLM Machine Translation for Macedonian

Jana Kuzmanova¹[0009-0009-5015-4183],
Sonja Gievska¹[0000-0002-3411-2399],
Katerina Zdravkova¹[0000-0002-9674-3081]

¹ Ss. Cyril and Methodius University in Skopje,
Faculty of Computer Science and Engineering, North Macedonia,
{jana.kuzmanova|sonja.gievska|katerina.zdravkova}@finki.ukim.mk

Abstract

Despite the increasing popularity of large language models for machine translation and their good performance across many language pairs, there is still considerable room for improvement, particularly for low-resource languages and translation into these languages. In-context learning approaches have shown promise in enhancing translation quality in such settings. Providing translations of words from the source text, similar parallel sentences, and grammatical information has resulted in improvements across multiple language pairs. In this paper, we explore whether and how different in-context learning approaches, including parallel sentences and dictionary entries, can improve translation from English into Macedonian. We examine different numbers of examples and different degrees of similarity to the source sentences, phrases, and words to determine which configurations are most effective. Our results show that the effectiveness of in-context learning depends strongly on the model, prompt configuration, and type of additional information provided. Greater example similarity does not consistently improve translation quality, while dictionary-based prompting generally yields poorer results, particularly when out-of-domain entries are used. Although improvements are observed for individual models and translations, all evaluated LLMs perform below the encoder-decoder NLLB baseline across all experimental settings.

Keywords: machine translation, large language models, in-context learning, Macedonian, low-resource languages

Istraživanje učenja u kontekstu za mašinsko prevođenje na makedonski jezik pomoću velikih jezičkih modela

Jana Kuzmanova, Sonja Gievska, Katerina Zdravkova

Sažetak

Uprkos sve većoj popularnosti velikih jezičkih modela u mašinskom prevođenju i njihovim dobrim rezultatima na brojnim jezičkim parovima, i dalje postoji značajan prostor za poboljšanje, posebno kada je reč o jezicima sa ograničenim resursima i prevođenju na te jezike. Pristupi zasnovani na učenju u kontekstu pokazali su potencijal za unapređenje kvaliteta prevoda u takvim uslovima. Uključivanje prevoda reči iz izvornog teksta, sličnih paralelnih rečenica i gramatičkih informacija dovelo je do poboljšanja rezultata na više jezičkih parova. U ovom radu istražujemo da li i na koji način različiti pristupi učenja u kontekstu, uključujući paralelne rečenice i rečničke odrednice, mogu poboljšati prevođenje sa engleskog na makedonski jezik. Ispitali smo različit broj primera i različite stepene njihove sličnosti sa izvornim rečenicama, frazama i rečima kako bismo utvrdili koje su konfiguracije najefikasnije. Naši rezultati pokazuju da efikasnost učenja u kontekstu u velikoj meri zavisi od modela, konfiguracije upita i vrste dodatnih informacija. Veća sličnost primera ne dovodi dosledno do boljeg kvaliteta prevoda, dok upiti zasnovani na rečničkim odrednicama uglavnom daju slabije rezultate, naročito kada se koriste odrednice iz drugog domena. Iako su kod pojedinačnih modela i prevoda zabeležena poboljšanja, svi testirani veliki jezički modeli postigli su rezultate ispod NLLB referentnog sistema u svim eksperimentalnim postavkama.

Ključne reči: mašinski prevod, veliki jezički modeli, učenja u kontekstu, jezici sa ograničenim resursima

Towards Fine-Grained Annotation of Directed Hate and Abusive Speech in Serbian

Miloš Košprdić¹[0000-0001-6902-3639], Isidora Dekić¹[0009-0005-6247-7669],
Tijana Radović¹[0000-0002-5759-0844]

¹University of Belgrade, Serbia,
{milos.kosprdic|isidora.r.dekic|tijana.n.radovic}@gmail.com

Abstract

Hate-speech detection is challenging because abusive language, profanity, identity-based derogation, and group-directed hostility frequently overlap while remaining conceptually distinct. This distinction is particularly relevant in Serbian online discourse, where offensive and identity-related expressions may perform different pragmatic functions without necessarily constituting hate speech. We present a fine-grained annotation framework and accompanying guidelines for Serbian abusive and hate speech that explicitly model target structure, identity-based derogation, profanity function, quotation and reported abuse, and author stance. A 133,483-comment Serbian Reddit candidate corpus was constructed through phenomenon-oriented lexical retrieval and preprocessing, from which an intentionally enriched 500-comment pilot dataset was selected. The pilot was first processed through a structured LLM annotation pipeline and then annotated independently by multiple human annotators, enabling analysis of annotation reliability, construction of a human-consensus reference, and direct evaluation of the LLM first pass. The results support a clear distinction between abusive speech, identity-based derogation, and hate speech, while showing that the presence of profanity alone has little discriminatory value. Pragmatic and target-related judgments proved more difficult than those concerning explicit lexical phenomena, and the LLM showed a systematic tendency to apply the hate-speech label too broadly to abusive and identity-related content. The study therefore provides an operational annotation framework, accompanying guidelines, and a human-validated Serbian pilot resource, while establishing an empirical basis for refining LLM-assisted annotation and scaling resource construction to larger Serbian-language collections.

Keywords: Hate Speech, Abusive Language, Identity-Based Derogation, Fine-Grained Annotation, Serbian NLP, Human Annotation, LLM-Assisted Annotation

Warning: *This paper contains examples of potentially offensive content.*

Ka višedimenzionalnoj anotaciji usmerenog govora mržnje u srpskom jeziku

Miloš Košprdić, Isidora Dekić, Tijana Radović

Sažetak

Detekcija govora mržnje predstavlja složen zadatak jer se uvredljivi govor (engl. abusive speech), psovke, omalovažavanje zasnovano na identitetu i hostilnost usmerena prema društvenim grupama često prepliću, iako ne predstavljaju iste komunikativne pojave. Ovo razgraničenje posebno je značajno u srpskom onlajn diskursu, u kojem uvredljivi i identitetski obojeni izrazi mogu imati različite pragmatičke funkcije i ne moraju nužno predstavljati govor mržnje. U radu predstavljamo višedimenzionalnu anotacionu šemu i prateće smernice za govor mržnje i uvredljivi govor u srpskom jeziku, kojima se eksplicitno modeluju struktura i vrsta mete, omalovažavanje zasnovano na identitetu, funkcija psovki, citirani i preneseni uvredljivi govor i autorski stav. Primenom leksički i fenomenološki usmerenog prikupljanja konstruisan je veći korpus komentara sa platforme Reddit, iz kojeg je izdvojen namerno obogaćen pilot skup od 500 komentara. Pilot skup je najpre anotiran strukturisanim postupkom zasnovanim na velikom jezičkom modelu, a zatim nezavisno od strane više ljudskih anotatora, što je omogućilo procenu pouzdanosti anotacione šeme, formiranje ljudskog konsenzusa i neposredno poređenje odluka modela i ljudskih anotatora. Rezultati potvrđuju jasnu razliku između uvredljivog govora, omalovažavanja zasnovanog na identitetu i govora mržnje, dok samo prisustvo psovki ima malu diskriminativnu vrednost. Pragmatičke odluke povezane sa funkcijom psovki, tumačenjem mete i autorskim stavom pokazale su se zahtevnijim od prepoznavanja eksplicitnih jezičkih pojava, dok je veliki jezički model pokazao sistematsku sklonost ka preširokom dodeljivanju oznake govora mržnje uvredljivom i identitetski obojenom sadržaju. Rad pruža operacionalizovanu anotacionu šemu, prateće smernice i ljudski validiran pilot resurs za srpski jezik, kao i empirijsku osnovu za unapređenje anotacije podržane velikim jezičkim modelima i njeno dalje skaliranje na veće kolekcije podataka.

Ključne reči: govor mržnje, uvredljivi govor, višedimenzionalna anotacija, obrada prirodnog jezika, srpski jezik, ljudska anotacija, anotacija podržana velikim jezičkim modelima

Translation Direction in the *Gesta regum Sclavorum*: A Quantitative and Qualitative Approach

Dragana Kunčer¹[\[0000-0002-4688-4775\]](#)

¹ Institute of History, Belgrade, Serbia, dragana.kuncer@iib.ac.rs

Abstract

The paper examines the claim that the Latin chronicle *Gesta regum Sclavorum*, a text concerning the medieval history of the South Slavs, was not originally composed in Latin but was translated from Italian into Latin as part of a deliberate forgery. The question of translation direction between the two versions is approached through quantitative, philological, and pragmatic analyses. The quantitative analysis is based on concepts such as translation universals, laws of translation, and translationese, according to which translated texts may be distinguished from texts originally composed in the target language by certain linguistic features. The hypotheses concerning the distinctive features of translated texts were operationalised for quantitative analysis. The analysis was performed on preprocessed versions of the two texts using Python scripts generated with the assistance of ChatGPT from prompts designed for this purpose. The philological analysis was conducted through a manual comparison of the parallel texts. The pragmatic analysis was based on Ivan Lučić's comments on the content of the chronicle, as expressed both in his historiographical work *De regno Dalmatiae et Croatiae libri sex* and in letters to his friends. Lučić was the first editor of the Latin text and, according to the theory under examination, is presumed to have been its translator and one of the two figures responsible for the alleged forgery.

Keywords: translation direction, translation universals, translationese, *Gesta regum Sclavorum*, spaCy, LatinCy, Python scripts

Смер превода *Gesta regum Sclavorum*: квантитативни и квалитативни приступ

Драгана Кунчер

Сажетак

У раду се испитује теорија Соланж Бујан, према којој латинска хроника *Gesta regum Sclavorum* (GRS), дело од значаја за средњовековну историју Јужних Словена, није изворно написана на латинском. Бујан сматра да хроника представља превод с италијанског на латински, објављен 1666. године у делу историчара Ивана Лучића, као и другу фазу кривотворења текста. Прва фаза била би састављање хронике на италијанском, објављене 1601. године од стране историчара Мавра Орбина. У раду се, међутим, полази од тога да је за проверу ове теорије довољно испитати хипотезу која се тиче друге фазе претпостављеног кривотворења, односно одредити смер превода између две верзије: уколико анализа потврди да је латински текст оригинал, а не превод, а да је, напротив, италијански текст превод, а не оригинал, самим тим хипотеза о првој фази, односно састављању хронике на италијанском, постаје неодржива. Како би се дошло до могућег решења проблема, урађена је квантитативна, филолошка и прагматичка анализа.

Квантитативна анализа полази од појмова као што су универзалије преводјења, закони преводјења и преводилачки језик (*translationese*), односно од претпоставке да се преведени текстови по појединим језичким особинама разликују од оригинално написаних текстова. Теоријске претпоставке о особинама преведеног текста операционализоване су тако да могу бити испитане квантитативно. На основу прегледа сличних истраживања издвојене су оне квантитативне мере које су се показале као могући показатељи преведених текстова: 1) лексичка разноврсност у преводу мања је од оне у оригинално написаном тексту; 2) просечна дужина реченица у преведеном тексту већа је од оне у оригинално написаном тексту; 3) лексичка густина, односно однос пунозначних и функцијских речи, у преведеном тексту мања је од оне у оригиналном тексту; 4) однос личних имена и заменица које се односе на лица већи је у преведеном тексту од оног у оригиналном тексту.

Текстови су претходно припремљени за рачунарску обраду, а мерења су извршена помоћу Python скрипти које су за италијански

текст користиле библиотеку spaCy, а за латински текст језичке моделе LatinCy. Скрипте су написане уз помоћ ChatGPT-a, на основу специјално дизајнираних промптова. Резултати прве три мере иду у прилог традиционалној теорији. Што се тиче четврте мере, разлика у односу личних имена и заменица које се односе на лица веома је мала, па њено тумачење захтева опрезност.

Филолошка анализа упоредних места у два текста показује, између осталог, да латински текст на више места садржи додатни материјал који се не може реконструисати на основу италијанске верзије. То се тумачи као последица поједностављивања у преводу с латинског на италијански, у којем су та места изостављена. С друге стране, примећено је постојање додатног материјала и у италијанској верзији, који се не може реконструисати на основу латинске верзије. Иако се ова појава, на први поглед, може на сличан начин тумачити као аргумент у прилог теорији Бујан, то јест као последица скраћивања италијанског текста у преводу на латински, она се ипак тумачи као аргумент у прилог традиционалној теорији. Наиме, сматра се да је тај материјал унет због тога што је Орбин за превод на италијански користио више латинских рукописа и да се налазио у рукопису који Лучић није познавао. Филолошка анализа такође показује да се у италијанској верзији јављају уопштавања, појашњавања породичних односа, синтаксичке и семантичке необичности, као и игра речи која је разумљива само на латинском. Ове појаве се најлакше објашњавају као последице превода с латинског на италијански.

Прагматичка анализа Лучићевих коментара показује да је он отворено доводио у питање историјску поузданост хронике, што је у нескладу са улогом кривотворитеља која му је приписана у теорији Бујан, те се може тумачити као аргумент у прилог традиционалној теорији.

Посматрани заједно, резултати квантитативне, филолошке и прагматичке анализе одговарају схватању да је италијанска верзија превод с латинског, а не обратно, чиме се додатно оснажује теза о средњовековном пореклу латинске верзије *Gesta regum Sclavorum*, а самим тим и о њеној аутентичности.

Кључне речи: смер превода, универзалије превођења, преводилачки језик, *Gesta regum Sclavorum*, spaCy, LatinCy, Python скрипте

Semantic Role Labeling

Modeling Expression of Eventuality Types in Bulgarian with Large Language Models

Svetla Koeva¹[\[0000-0001-5947-8736\]](#),

¹ Bulgarian Academy of Sciences, Institute for Bulgarian Language, Bulgaria,
svetla@dcl.bas.bg

Abstract

The paper examines the relationship between Bulgarian verbal aspect and eventuality types. It adopts a semantic framework based on two privative oppositions, intervality and terminativity, to distinguish eventuality types, and applies these features to Bulgarian perfective and imperfective verbs. The proposed features provide a systematic account of how Bulgarian imperfective and perfective verbs are distributed across eventuality types, while also capturing contextual coercion. A dataset of 254 Bulgarian verbs (the *Aspect:BG* dataset) was annotated with appropriate derivational and semantic features for each verb. Multiple experimental runs investigated the use of Large Language Models (LLMs) to evaluate human classification and to further support semi-automated classification. The analysis shows that different LLMs can be used individually or in combination to generate annotations for an expanded version of the dataset, with different models applied to different semantic or derivational features; however, expert verification is still recommended.

Keywords: eventuality types, Bulgarian verb aspect, Large Language Models

Modelovanje izražavanja tipova eventualnosti u bugarskom jeziku pomoću velikih jezičkih modela

Svetla Koeva

Sažetak

Rad ispituje odnos između bugarskog verbalnog aspekta i tipova eventualnosti. Usvaja semantički okvir zasnovan na dve privativne opozicije, intervalnosti i terminativnosti, kako bi se razlikovali tipovi eventualnosti, i primenjuje ove karakteristike na bugarske perfektivne i imperfektivne glagole. Predložene karakteristike pružaju sistematski prikaz načina na koji se bugarski imperfektivni i perfektivni glagoli raspoređuju po tipovima eventualnosti, a istovremeno obuhvataju kontekstualnu koerciju. Skup podataka od 254 bugarska glagola (skup podataka *Aspect:BG*) anotiran je odgovarajućim derivacionim i semantičkim karakteristikama za svaki glagol. Upotreba velikih jezičkih modela (LLM) za procenu ljudske klasifikacije i za podršku poluautomatskoj klasifikaciji i proširenju skupa podataka istražena je u nekoliko eksperimenata. Analiza pokazuje da se različiti LLM-ovi mogu koristiti pojedinačno ili u kombinaciji za generisanje anotacija za proširenu verziju skupa podataka, pri čemu se različiti modeli primenjuju na različite semantičke ili derivacione karakteristike; međutim, i dalje se preporučuje stručna verifikacija.

Ključne reči: tipovi situacija, bugarski glagolski vid, veliki jezički modeli

Approaches to Automatic Annotation of Semantic Roles and Syntactic Relations

Svetlozara Leseva¹[\[0000-0001-8198-4555\]](#), Ivelina
Stoyanova¹[\[0000-0003-3771-435X\]](#)

¹ Bulgarian Academy of Sciences, Institute for Bulgarian Language, Bulgaria,
{zarka|iva}@dcl.bas.bg

Abstract

We investigate the capabilities of several openly available large language models (LLMs) for the automatic annotation of frame-semantic structure in Bulgarian, a morphologically rich, pro-drop Slavic language for which the application of LLMs to FrameNet-style annotation remains largely unexplored. The study includes a controlled pilot annotation on 350 sentences containing cleaning verbs (e.g. *чистя, мета, мия, бърша*) that evoke two closely related FrameNet frames, **Removing** and **Emptying**, which share the same core elements – AGENT, THEME, SOURCE, but differ in their syntactic profiling which distinguish frame-semantic alternations. We evaluate three models – Gemma 4 E4B-IT, BgGPT-Gemma-3-12B-IT and Qwen3-8B, on frame element assignment, with the additional evaluation of four annotation subtasks (clause-boundary identification, phrase span identification, coreference resolution, pro-drop resolution). The Bulgarian-adapted BgGPT-Gemma-3-12B-IT achieves the highest scores on all five tasks (F_1 of 58.4), followed by Qwen3-8B with chain-of-thought reasoning (50.9) and Gemma 4 E4B-IT (38.7). Bulgarian fine-tuning matters most where morphosyntax is decisive – pro-drop, clause segmentation, phrase boundaries, while reasoning partially compensates on the more conceptual tasks. The results motivate further work on retrieval-based prompt construction, iterative self-refinement, and supervised fine-tuning.

Keywords: FrameNet, frame elements, semantic annotation, large language models

Pristupi automatskom označavanju semantičkih uloga i sintaksičkih odnosa

Svetlozara Leseva, Ivelina Stojanova

Sažetak

U radu se ispituju mogućnosti nekoliko javno dostupnih velikih jezičkih modela (LLM) za automatsko obeležavanje semantičkih uloga u bugarskom, morfološki bogatom, slovenskom pro-drop jeziku. Primena velikih jezičkih modela za ovu vrstu anotacije u bugarskom i dalje je nedovoljno istražena. Studija obuhvata kontrolisanu pilotsku anotaciju 350 rečenica koje sadrže glagole čišćenja (npr. чистя, мета, мия, бърша). Ovi glagoli prizivaju dva blisko povezana okvira u FrameNet-u, **Uklanjanje** (Removing) i **Pražnjenje** (Emptying), koji dele iste centralne elemente – AGENS, TEMU i IZVOR – ali ih različito sintaksički profilisu, na osnovu čega se uočavaju alternacije okvira. U radu se ocenjuju tri modela — Gemma 4 E4B-IT, BgGPT-Gemma-3-12B-IT i Qwen3-8B — u zadatku pripisivanja elemenata okvira, uz dodatnu evaluaciju za četiri podzadatka anotacije: određivanju granica klauza, identifikovanju raspona sintagmi, razrešavanju koreferencije i određivanju izostavljenog subjekta. Model BgGPT-Gemma-3-12B-IT, prilagođen bugarskom jeziku, postiže najbolje rezultate u svih pet zadataka ($F1 = 58,4$), a slede ga Qwen3-8B uz lančano zaključivanje (50,9) i Gemma 4 E4B-IT (38,7). Fino podešavanje (*fine-tuning*) za bugarski jezik najznačajnije je tamo gde je morfosintaksa presudna – kod izostavljanja zamišljenog subjekta (pro-drop), segmentacije rečenica i granica sintagmi – dok sposobnost zaključivanja delimično kompenzuje nedostatke na zadacima koji su više konceptualne prirode. Dobijeni rezultati daju podsticaj za dalji rad na kreiranju upita zasnovanim na pretraživanju (*retrieval-based*), iterativnom samousavršavanju i nadgledanom finom podešavanju (*supervised fine-tuning*).

Ključne reči: FrameNet, elementi okvira, semantička anotacija, veliki jezički modeli

A FrameNet Analysis of Serbian Verbs *Kidati* and *Cepati*: Frames vs Traditional Word Sense Disambiguation

Marina Bagi¹[\[0009-0006-9410-7862\]](#), Aleksandra Marković¹[\[0000-0003-2985-3042\]](#),
Ely Matos²[\[0000-0002-9464-9315\]](#)

¹ Serbian Academy of Sciences and Arts, Institute for the Serbian Language,
Serbia, {aleksandra.markovic|marina.bagi}@isj.sanu.ac.rs

²Universidade Federal de Juiz de Fora, Brasil, ely.matos@ufjf.br

Abstract

This paper presents an analysis of two Serbian verbs, *kidati* (‘tear’, ‘rip’, depending on context) and *cepati* (‘tear’, ‘rip’, ‘chop’), from a Frame Semantics perspective and compares the results of this analysis with the senses of these verbs identified in two Serbian monolingual descriptive dictionaries. This research has multiple aims: to offer a Frame Semantics analysis of these verbs; to populate the *Serbian FrameNet* database with new frames evoked by these verbs; to broaden the scope of the existing frame inventory; and to highlight new insights from this analysis and compare them with insights from traditional lexicographic description. The analysis, conducted on around 180 sentences containing these two verbs from three Serbian language corpora, follows a corpus-driven, lemma-by-lemma approach to examine all senses of the target verbs. In addition, we adopted the lexicographic annotation of corpus sentences selected for predetermined target lexical units.

Our analysis yielded 21 frames for the verb *kidati* and 18 frames for the verb *cepati*. Because these verbs are partial synonyms, we identified 9 frames shared by both verbs. The analysis was driven by the semantic and syntactic distribution of lexical units (sublexemes), not by the predetermined decision on the number of semantic frames. Thus, we reached the limits of the existing frame inventory and had to introduce new frames. Finally, we compared our results with the descriptions and sense inventories of these lexemes in Serbian descriptive dictionaries. This comparison showed, as expected, that frames and dictionary senses do not correspond one-to-one. Instead, they are related in a complex way.

Keywords: *Serbian FrameNet*, Frame Semantics, verb *kidati*, verb *cepati*, lexicography

Analiza glagola *kidati* i *cepati* u Frejmnetu: okviri naspram tradicionalnog izdvajanja značenja reči

Marina Bagi, Aleksandra Marković, Eli Matos

Sažetak

Ovaj rad predstavlja analizu dva srpska glagola, *kidati* i *cepati*, iz ugla teorije semantike okvira (Frame Semantics) i upoređuje rezultate ove analize sa značenjima ovih glagola identifikovanim u dva srpska jednojezična opisna rečnika. Istraživanje ima nekoliko ciljeva: da ponudi analizu ovih glagola zasnovanu na semantici okvira; da dopuni bazu podataka *Serbian FrameNet* novim okvirima koje ovi glagoli prizivaju; da proširi postojeći inventar okvira; i da istakne nove uvide proizašle iz ove analize i uporedi ih sa uvidima iz tradicionalnog leksikografskog opisa. Analiza sprovedena na oko 180 rečenica koje sadrže ova dva glagola iz tri korpusa srpskog jezika prati pristup vođen korpusom (*corpus-driven*), kao i tzv. lema po lema pristup (*lemma-by-lemma*), kako bi se ispitala sva značenja ciljnih glagola. Pored toga, primenili smo leksikografsku anotaciju korpusnih rečenica izabranih za unapred određene ciljne leksičke jedinice.

Kao rezultat analize izdvojen je 21 okvir za glagol *kidati* i 18 okvira za glagol *cepati*. Pošto su ovi glagoli delimični sinonimi, identifikovali smo 9 zajedničkih okvira za oba glagola. Analiza je bila vođena semantičkom i sintaksičkom distribucijom leksičkih jedinica (subleksema), a ne unapred donetom odlukom o broju semantičkih okvira. Zbog toga postojeći inventar okvira nije bio dovoljan i morali smo da uvedemo nove okvire. Na kraju, uporedili smo naše rezultate sa njihovim značenjima i definicijama u srpskim opisnim rečnicima. Ovo poređenje je pokazalo, kao što se i očekivalo, da između okvira i rečničkih značenja ne postoji poklapanje jedan prema jedan. Umesto toga, oni su povezani na složen način.

Ključne reči: *Serbian FrameNet*, semantika okvira, glagol *kidati*, glagol *cepati*, leksikografija



Društvo za jezičke resurse i tehnologije – JeRTeh
Друштво за језичке ресурсе и технологије – JePTex
Language Resources and Technologies Society – JeRTeh

<https://judig.jerteh.rs/>



Vektorizacija teksta – aplikacije za srpski jezik
Векторизација текста – апликације за српски језик
Text Embeddings – Serbian Language Applications

<https://tesla.rgf.bg.ac.rs/>