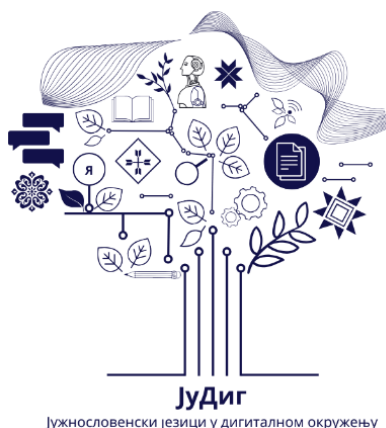


JuDig





**Proceedings of the International Conference
South Slavic Languages in the Digital Environment
JuDig
Thematic Collection of Papers**

Accepted by the Teaching and Scientific Council of the Faculty of Philology, University of Belgrade, on 27 November 2025, based on the positive reviews Dr. Aleksandra Marković, Senior Research Associate at the Institute for the Serbian Language SASA, and Dr. Milica Dinić Marinković, Assistant Professor at the Faculty of Philology, University of Belgrade.



ORGANISERS

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

Belgrade 2025

Edition:

Научни скупови
Proceedings of the International Conference South Slavic Languages in the
Digital Environment, JuDig, Thematic Collection of Papers

Publisher:

University of Belgrade – Faculty of Philology
Studentski Trg 3, Belgrade, Serbia

For publisher:

Prof. dr Iva Draškić Vićanović

Editors:

Prof. dr Jasmina Moskovljević Popović  0009-0003-5737-0184
University of Belgrade – Faculty of Philology
Prof. dr Ranka Stanković  0000-0001-5123-6273
University of Belgrade – Faculty of Mining and Geology
Language Resources and Technologies Society JeRTeh

Technical editors:

Dr Aleksandra Tomašević
Prof. dr Ranka Stanković

Graphic design:

Dr Aleksandra Tomašević

Published by: MAB, Belgrade

Circulation: 10 copies

Publication accessible at:

<https://judig.jerteh.rs/2024/judig-2024.pdf>

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

E-mail: judig@jerteh.rs

ISBN: 978-86-6153-791-2

DOI: 10.18485/judig.2025.1

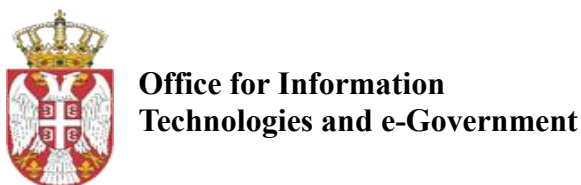
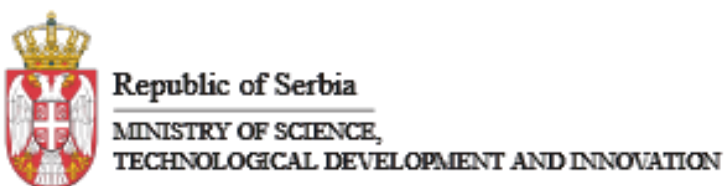


<https://creativecommons.org/licenses/by/4.0/>
Works are licensed under the terms and conditions of the
Creative Commons Attribution 4.0 International License.

The conference is supported by



(bronze sponsor)



PROGRAMME COMMITTEE AND REVIEWING TEAM

Co-Chairs:

Dr Jasmina Moskovljević Popović, full professor,
University of Belgrade – Faculty of Philology

Dr Ranka Stanković, full professor
University of Belgrade – Faculty of Mining and Geology, and
Language Resources and Technologies Society JeRTeh

Members:

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

Dr Aleksandra Marković, Institute for Serbian, SANU, Serbia

Dr Ana Ostroški Anić, Institute of Croatian Language, Croatia

Doc. dr Baša Stipčević, University of Belgrade, Faculty of Philology,
Serbia

Prof. dr Benedikt Perak, University of Rijeka, Faculty of Philosophy,
Croatia

Dr Biljana Rujević, University of Belgrade, Faculty of Mining and
Geology, Serbia

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

Prof. Vera Čevriz Nišić, University of East Sarajevo, Faculty of
Philosophy, Bosnia and Herzegovina

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

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

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

Prof. Gordana Pavlović-Lažetić, Language Resources and Technologies
Society (JeRTeh), Serbia

Ass. Prof. Danilo Aleksić, University of Belgrade, Faculty of Philology,
Serbia

Prof. Dimitar Trajanov, Saints Cyril and Methodius University - Skopje,
Faculty of Computer Science and Engineering, Macedonia

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

Dr Jaka Čibej, University of Ljubljana, Faculty of Philosophy, Slovenia

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

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

Prof. Jelena Marković, University of East Sarajevo, Faculty of
Philosophy, Bosnia and Herzegovina

Doc. Jovan Čudomirović, University of Belgrade - Faculty of Philology,
Serbia

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

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

Ass. Prof. Miloš Utvić, University of Belgrade - Faculty of Philology,
Serbia

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

Dr Nebojša Vasiljević, Petlja Foundation, Serbia

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

Prof. Nelda Kote, Polytechnic University of Tirana, Albania

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

Dr Paraskevi Giouli, Institute for Language and Speech Processing,
Athens Research Center, Greece

Prof. Petja Osenova, Sofia University "St. Kliment Ohridski", Bulgaria

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

Ass. Prof. Saša Marjanović, University of Belgrade - Faculty of
Philology, Serbia

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

Prof. Svetla Koeva, Bulgarian Academy of Sciences, Bulgaria

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

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

Dr Kristina Štrkalj Despot, Institute of Croatian Language, Croatia

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

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

ORGANISING COMMITTEE

Co-Chairs:

Dr Jasmina Moskovljević Popović, full professor,
University of Belgrade – Faculty of Philology

Dr Ranka Stanković, full professor
University of Belgrade – Faculty of Mining and Geology, and
Language Resources and Technologies Society JeRTeh

Members:

Ass. Prof. Jovan Čudomirović, Faculty of Philology, University of
Belgrade

Ass. Prof. Balša Stipčević, Faculty of Philology, University of Belgrade

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

Ass. Prof. Miloš Utvić, Faculty of Philology, University of Belgrade

Ass. Prof. Milica Dinić Marinković, Faculty of Philology, University of
Belgrade

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

Dr Biljana Rujević, Language Resources and Technologies Society
(JeRTeh)

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

Dr Aleksandra Marković, Language Resources and Technologies Society
(JeRTeh)

Anđelka Zečević, Language Resources and Technologies Society
(JeRTeh)

Nikola Gudžić, Language Resources and Technologies Society (JeRTeh)

Table of Contents

Editors' Preface	1

<i>Душко Витас, Ранка Станковић, Цветана Крстев</i> О фамилији корпуса савременог српског језика СрпКор	3

<i>Jaka Čibej</i> Models for Automatic Morphological Inflection of Serbian and Croatian Based on the srLex and hrLex Morphological Lexicons	29

<i>Jaka Čibej, Ranka Stanković, Ana Ostroški Anić, Simon Krek, Carole Tiberius</i> The ELEXIS-WSD Parallel Sense-Annotated Corpus and South Slavic Languages: Subcorpora for Croatian, Serbian, and Slovene	45

<i>Jelena Redli</i> Значај digitalnog korpusa i jezičkih alata u jezičkoj analizi forenzičkih tekstova na srpskom jeziku.....	61

<i>Katerina Zdravkova, Jana Kuzmanova</i> Sonority Based Syllabic Hyphenation of Macedonian and Serbian.....	77

<i>Јелена Марковић</i> Ученички корпуси србофоних говорника енглеског језика у контрастивној анализи међујезика на примеру језичке неодређености	89

<i>Maram Alharbi, Ruslan Mitkov</i> Deep Learning versus Rule-based Models: Is Rule-based Methodology a Thing of the Past? A Study of Hotel Review Sentiment Analysis	113

<i>Ivelina Stoyanova, Svetlozara Leseva</i> Towards Classifying Predicates Involving Change	129

Ana Kovačević

Disinformation and Generative AI: Risks, Challenges,
and Possible Solutions 147

Jelena Lazarević, Olivera Kitanović

Contrastive Analysis of Syntax Patterns in Comparable
Football Corpora in Spanish and Serbian Languages 163

Марина Багу

Синтаксичка и семантичка анализа глагола оштетити и
уништити из угла теорије семантике оквира 181

Срђан Шућур, Јелена Марковић

Дигитализација српског књижевног наслеђа ијекавског
изговора (1840–1920) при Центру за дигиталну хуманистику
Филозофског факултета Пале (прва фаза) 209

*Снежана Петровић, Мирјана Петровић Савић, Ана Шпановић,
Ленка Бајчетић, Матија Нешовић, Јована Тодорић*

Дигитализација грађе Одбора за ономастику САНУ –
значај, циљеви и први кораци 229

Ranka Stanković, Jovana Rađenović, Maja Ristić, Dragan Stankov

Creation of a Training Dataset for Question-Answering Models
in Serbian..... 245

Saša Moderc

Challenges and Perspectives in Italian Clitic Tagging:
A Case Study on the SerbltaCor3 Corpus 265

Olja Perišić

Korpusi za učenje srpskog jezika kao stranog u eri veštačke
inteligencije 279

Marija Đokić Petrović, Mihailo St. Popović, Vladimir Polomac

Named Entity Recognition for Pre-modern Serbian:
A Preliminary Study 295

Martina Pavić

Frekvencija značenja pridjeva u hrvatskome medicinskom nazivlju –
korpusno utemeljena analiza kolokacija..... 311

Mihailo Škorić, Saša Petalinkar

Quality Textual Corpora and New South Slavic Language Models 337

Nataša Kiš

Sintaksičko-semantička anotacija elektronskog korpusa
srpskog jezika..... 349

Kristina Ilić

Značaj paralelnih korpusa za istraživanje frazemskih konstrukcija
u nemačkom i srpskom jeziku 363

Данило Алексић

Нове унапређене верзије програма „Ка минималним паровима” 381

Danka Jokić, Ranka Stanković

Exploring the Synergy Between LLMs and Knowledge
Graphs for Advanced Abusive Speech Detection in Serbian 395

Nikitas N. Karanikolas

Noun Phrase and Prepositional Phrase Chunking for the Greek
Language with spaCy..... 411

Dušanka Vujović, Branko Milosavljević

Korpus Srpsko kao tehnološka osnova za izradu Rečnika
savremenog srpskog jezika Matice srpske 425

Nebojša Ratković

Integration of Serbian Language Version of Wikipedia into Educational
Systems and the Advancement of Language Technologies 441

Никола Јанковић, Јована Иваниш

Употреба модела Whisper Large v3 Sr за транскрипцију говора
на српском језику у програмском језику Пајтон на платформи
Гугл колаб..... 453

Maja Matijević

Od rječnika kao korpusa do hiponimije i meronimije 473

Nikola Janković

Metodologija izrade višjezičnog paralelnog korpusa na osnovu
onlajn digitalnih uputstava za upotrebu: korpus Hilti uputstava 491

Editors' Preface

The international scientific conference South Slavic Languages in Digital Environment—JuDig, which took place in Belgrade from 21 to 23 November 2024 and was organised by the Faculty of Philology, University of Belgrade and Language Resources and Technologies Society (JeRTeh), provided an opportunity for numerous researchers in the field of computational linguistics, language technologies, and related fields to share their ideas, insights, and results in these research areas, as well as to discuss the latest advancements and innovations in the field. The main aim of the conference was to foster networking and to provide a forum for high-quality research in all (sub)domains of computational linguistics and related topics, with a special focus on language resources and technologies available for Serbian and other South Slavic and Balkan languages. Forty-three eminent scientists from twelve countries and twenty-two research institutions served on the Conference Programme Committee.

Fifty-five contributions on a wide range of topics in language technologies were submitted to the conference. All submissions underwent a double-blind peer review process by three members of the Programme Committee. The Programme Committee and the co-chairs selected 51 research papers written by 84 (co)authors from 34 institutions and 14 countries to be presented at the conference – fifty-five from Serbia, five from Croatia, four from Bulgaria and Germany, two from Slovenia, Austria, Macedonia, United Kingdom, France, and Bosnia and Herzegovina, and one from Greece, the Netherlands, South Korea, and Italy. The Conference was held in a hybrid format and included four invited talks, 12 parallel sessions, and 4 thematically oriented workshops.

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

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

The proceedings present 29 contributions on various topics by researchers from Serbia, Slovenia, Croatia, Bosnia and Herzegovina, Northern Macedonia, United Kingdom, Greece, Italy, Austria, Nederland, Bulgaria, and South Korea. All submissions underwent a double-blind peer review process by the members of the Editorial Board. The publication serves as a comprehensive record of the conference's valuable insights, findings, and discussions, providing a relevant resource for professionals and scholars in the field.

We hope this collection of papers will inspire further research, collaboration, and progress in this dynamic, ever-evolving field. We thank all the contributors and attendees who made this conference a success, and we look forward to continuing the conversation and exploration of new ideas in the future. The preparation and publication of this volume were supported by the project 7276, Text Embeddings - Serbian Language Applications - TESLA, funded by the Science Fund of the Republic of Serbia.

Finally, we would like to thank everyone who participated in organizing the conference and preparing the proceedings. Our thanks also go to the sponsors and supporters of the Conference. Their generous contributions and commitment have been instrumental in making this event possible. We believe that with joint efforts, we could make a considerable impact on the future of research in this crucial scientific area.

Belgrade, October 2025
The Editors

О фамилији корпуса савременог српског језика СрпКор

Научни рад

DOI: 10.18485/judig.2025.1.ch1

Душко Витас,  0000-0003-4194-692X

Ранка Станковић,  0000-0001-5123-6273

Цветана Крстев¹  0000-0003-3328-9392

Апстракт

Рад приказује мотиве за развој и еволуцију СрпКор-а, описујући промене у коришћеним алатима, у димензијама корпуса, временској покривености, као и нивоима анотације. Поред грађе са веба, у овим корпусима је присутна и грађа која није доступна на мрежи, нарочито из научног и књижевног домена, али и она која због технолошких промена, није била архивирана на изворним сајтовима. Биће описан и његов даљи развој који укључује обogaћивање метаподатака, допуну анотација и унос нових садржаја осим материјала прикупљеног са веба. Актуелна верзија корпуса је повезана са другим језичким, посебно лексичким ресурсима, кроз систем Лексимирика што омогућава значајно унапређење лексикографског рада, али и оригинално интегрисано окружење за језичке експерименте.

Кључне речи: СрпКор, корпуси, српски језик, лематизација, Лексимирика

1. Увод

Акроним СрпКор означава фамилију електронских корпуса савременог српског језика чија је изградња почела крајем седамдесетих година прошлог века, а која је постала шире видљива заинтересованој истраживачкој заједници објављивањем његове прве верзије на вебу 2003. године. У овом дугом периоду, посебно пре појаве корисних текстуелних ресурса на вебу, развој корпуса се састојао у прикупљању и

¹ Друштво за језичке реурсе и технологије JePTex, {vitas|cvetana|ranka}@jerteh.rs

обради грађе као и у развоју метода обраде корпуса. Наиме, електронски корпус није само колекција текстова у дигиталном облику (како се то, на пример, наводи у (Добрић 2012)), већ подразумева више компонената које ће заједно овакву колекцију учинити корисном у језичким и другим истраживањима. Ове компоненте, поред самих текстова, чине, пре свега, софтверска подршка организацији и експлоатацији колекције текстова и средства за различите нивое анотације текстова који ће се наћи у корпусу (Витас 2023).

СрпКор је, водећи рачуна о овим компонентама, током своје изградње прошао различите метаморфозе које пружају слику, како о еволуцији софтверске подршке за конструкцију и експлоатацију корпуса, тако и о развоју система анотација на различитим нивоима (метаподаци, морфолошко обележавање, лематизација, именовани ентитети, итд).

Крајње скромни услови (у поређењу са другим срединама, како у броју истраживача укључених у изградњу корпуса, додељеним финансијским средствима из различитих извора, расположивом опремом) су наметнули стратегију поступног развоја корпуса која је подразумевала да ће се нове верзије корпуса ослањати на материјал припремљен и употребљен у оним верзијама које су јој претходиле.

У раду ће бити илустрована еволуција у развоју СрпКор-а почев од његове прве верзије до данас пратећи упливе различитих средстава која су коришћена у изградњи појединачних верзија, као и промене димензија и система анотације текстова. Посебно ће бити описана структура појединачних верзија корпуса, њихове димензије, обухваћени временски период и ниво анотације.

Основне замисли приликом конципирања веб-верзије корпуса су прво изложене у (Витас & Поповић 2003), а затим у (Utvić 2013) где су описани бројни детаљи за верзију СрКор-а из 2013. године. Интеракције корпуса са речницима су разматране у (Krstev & Vitas 2005, Vitas & Krstev 2012).

У одељку 2 су описани први кораци обраде природних језика у Србији. Одељак 3 је посвећен првим корпусима на вебу, док се и одељку 4 говори о даљем развоју корпуса и њиховим верзијама.

2. Први кораци

Интересовање за (рачунарску) обраду природних језика је у Србији почело почетком осамдесетих година прошлог века, знатно касније него у другим европским срединама, па чак и оним које су и језички и географски биле блиске у то доба. Могући узроци за овакво

занемаривање једне важне језичке технологије данас су тешко разумљиви. Могуће их је тражити, на пример, у традиционалном одсуству математичког и информатичког образовања филолога у Србији, чије се последице и даље живо осећају. Овоме је могао посебно допринети и одјек Пирсовог извештаја (Pierce и други 1966.) који је негативно оценио развој дорцтаунског пројекта аутоматског превођења и обуставио његово финансирање². Како је део овог пројекта био финансиран у Југославији (стр. 41 и даље), обустављање овог пројекта је морало изазвати код традиционално образованих лингвиста оправдано зазирање од рачунарских технологија, (видети, на пример, (Ivić, 2001, стр. 61 и даље), (Rajić, 1981)). Зазирање од информатичких поступака у обради језичких података се види и кроз читав низ различитих квантитативних језичких истраживања која су до осамдесетих година прошлог века обављена ручним пребројавањем и узбуђавањем одређеног језичког материјала (погледати (Витас, 2023)).

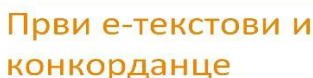
С друге стране, информатички оријентисани математичари у то исто доба се упознају са теоријом формалних језика и, посебно, са теоријом формалних граматика и језика Ноама Чомског и њиховом применом у компилацији програмских језика (Витас, 2018). Ова теорија природно буди и занимање за своју примену на природне језике, али су за такве експерименте недостајали неопходни ресурси – језичка грађа и програми за њену обраду. Ово је у извесном смислу последица одсуства интересовања филолога за обраду језика која је у другим срединама била покретач развоја потребних информатичких средстава.

Крајем 1977. године је у Сарајеву, у организацији Института за језик и књижевност, одржан скуп под насловом „Компјутерска обрада лингвистичких података“ (Šipka, 1978). Једини рад у коме се рачунар примењивао на материјала српског (српскохрватског) језика је приказивао неке од расподела графема у тексту (Томић, 1978). У другим срединама ондашње Југославије била су увелико у току истраживања везана за корпусе (пре свега у Загребу), као и различити програмски експерименти са словеначким језиком.

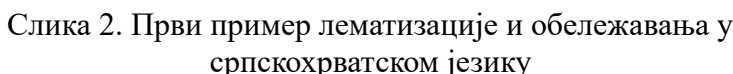
Подстакнуте овим скупом, започете су и у Београду активности на подручју обраде српског језика. Први кораци су се састојали у развоју неопходне програмске опреме. Тако је развијен властити систем за обраду корпуса, Аурора, који је улазни текст обрађивао у две фазе (Vitas 1979). У првој је формирана интерна репрезентација улаза у облику инвертованог текста, који није зависио од улазне азбуке³. У

² https://en.wikipedia.org/wiki/Georgetown%E2%80%93IBM_experiment

³ Текстови у систему АУРОРА су били кодирани користећи ознаке CX, CY,... за Ћ, Ч,... или Џ, Џ,... што је омогућило обраду која не зависи од азбуке изворног текста.



Непосредно за овим кораком се приступило сложенијем питању: формализацији именске морфологије, а резултат је био генератор именичке и придевске деklinације на основу морфографемске дефиниције леме (слика 2). Овај корак је омогућио израду парцијално лематизираних конкорданци текста, а представљао је и почетак рада на пројекту израде морфолошког електронског речника српског језика (Vitas 1980, Витас 2023).



У исто време, током 1978, формиран је Семинар за математичку и рачунарску лингвистику у Математичком институту⁴, а нешто касније, почев од 1981. године, и научно-истраживачки пројекат „Математичка и рачунарска лингвистика“ где су се окупили, с једне стране, математичари и информатичари, а са друге, лингвисти. Занимљиво је да су у том раном периоду живо интересовање за обраду језика показивали лингвисти коју се бавили другим језицима, а не српским (англисти, германисти, романисти и други). Крајем 1981. године, Математички институт је на Филолошком факултету организовао тродневни семинар „О репрезентативним узорцима природних језика (теорија корпуса)“ на коме је др Волфганг Тојберт са Института за немачки језик у Манхајму излагао немачка искуства у изградњи и експлоатацији корпуса.

Из ових почетних активности у Србији се да видети да је развој метода обраде српског језика почео међу информатичарима за разлику од других средина у којима су лингвисти били носиоци таквих активности. Природно је да су информатичари били заинтересовани, пре свега, за информатичке методе које омогућавају обраду једног језика, док су питања која се постављају једном лингвисти била у другом плану – информатичке компетенције их не могу формулисати.

Могућности масивних обрада текста су биле у оно доба вишеструко ограничене. С једне стране су то биле расположиве информатичке технологија где је унос текста вршен преко бушених картица, са врло ограниченим меморијским капацитетима (унутрашња меморија од 128KB) и користећи неприкладни програмски језик (Fortran IV). С друге стране, текстови у е-облику нису били доступни премда је већина великих штампарија у Београду већ тада користила затворене рачунарске системе са излазом на фотослогу. Из оваквих система није било могуће добити потребан излаз на магнетној траци или неком другом носачу који би омогућио преузимање већ унетих текстова у друге рачунарске системе. Такође, језици за обележавање текста, типа SGML и XML, који омогућавају преносивост, вишеструко коришћење и одрживост, још нису били развијени.

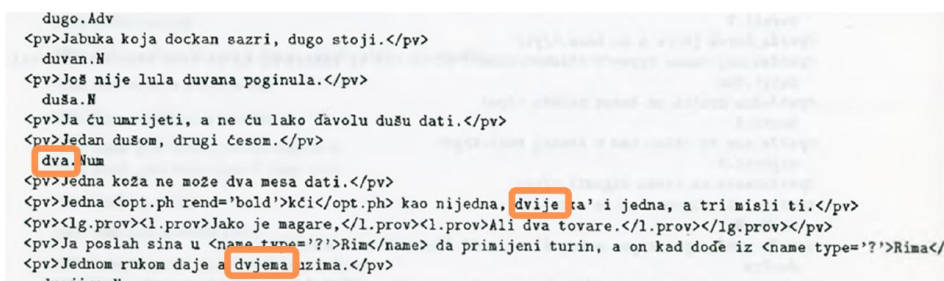
Упркос овим ограничењима, током прве деценије обраде српског језика било је прикупљено текстова у укупној дужини од око милион речи. Ова колекција се састојала од текстова уџбеника, литерарних и законских текстова и није имала посебно издиференцирану структуру. Значајан део литерарних текстова стигао је захваљујући Марк Хенингу

⁴ Далеки наследник овог семинара је данас семинар који се одржава у организацији Друштва за језичке ресурсе и технологије – JePTex.

са Универзитета у Орхусу који је сканирао извештај број романа на српскохрватском језику⁵.

Напредак у обради српског језика се огледа и у три конференције са међународним учешћем које је, као наставак сарајевске конференције, организовао институт Јожеф Штефан из Љубљане 1982, 1985. и 1988. године. Преглед ових зборника показује убрзани раст броја учесника из Србије, као и разноврсност тема које су обрађиване. Занимљиво је да се у опсежној библиографији радова из опште лингвистике за 1988. годину (Филолог 46⁶) забележен само један од радова са конференције одржане 1988. године: остали радови на материјалу српског језика нису забележени јер су долазили из информатичких средина!

Прва верзија морфолошког речника и идеја протографема помоћу система Аурора је имплементирана у оквиру докторске дисертације на тексту Вукових пословица (Krstev 1996, Krstev 1997). Протографема је коришћена на нивоу интерне репрезентације речника да омогући екстракцију различитих дијалекатских облика одреднице у аутоматској лематизацији (Слика 3).



```

dugo.Adv
<pv>Jabuka koja dockan sazri, dugo stoji.</pv>
duvan.N
<pv>Još nije lula duvana poginula.</pv>
duša.N
<pv>Ja ću umrijeti, a ne ću lako đavolu dušu dati.</pv>
<pv>Jedan dušom, drugi česom.</pv>
dva.Num
<pv>Jedna koža ne može dva mesa dati.</pv>
<pv>Jedna <opt.ph rend='bold'>kći</opt.ph> kao nijedna, dvije ka' i jedna, a tri misli ti.</pv>
<pv><lg.prov><l.prov>Jako je magare,</l.prov><l.prov>Ali dva toware.</l.prov></lg.prov></pv>
<pv>Ja poslah sina u <name type='?'>Rim</name> da primijeni turin, a on kad dođe iz <name type='?'>Rima</pv>
<pv>Jednom rukom daje a dvjema uzima.</pv>

```

Слика 3. Лематизирани конкорданце Вукових пословица

Од краја осамдесетих година прошлог века, истраживања српског језика су била укључена прво у кооперативни пројекат „Језичке индустрије“, а затим и у пројекат Telri⁷ (Trans-European Language Resources Infrastructure) средином деведесетих година двадесетог века упркос санкцијама Уједињених нација које су истраживачима отежавале међународну сарадњу. Ови пројекти су омогућили да се унапреде методе обраде српског језика кроз сарадњу са истраживачким групама других европских језика у овој области. Посебно значајни резултати су били морфосинтаксички опис српског језика у формализму Multext-East

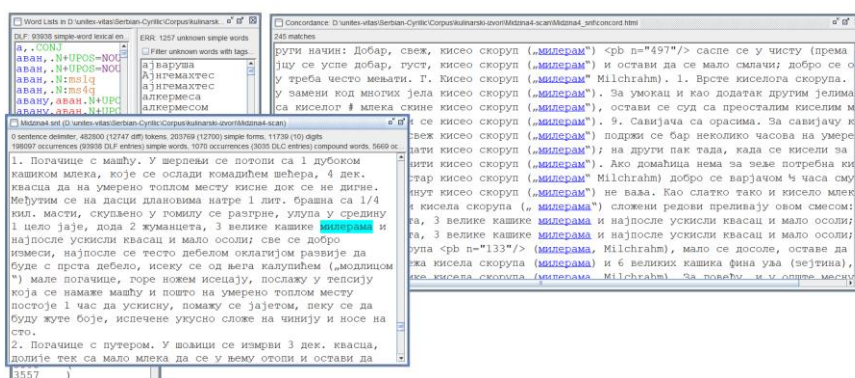
⁵ <http://www.yuope.com/books/yu/>

⁶ https://dais.sanu.ac.rs/bitstream/id/4224/bitstream_4224.pdf

⁷ <http://telri.nytud.hu/start.html> i <http://telri.nytud.hu/telriass/index.html>

(Krstev и други 2004) и примена овог описа у обележавању и лематизацији Орвеловог романа 1984 (Krstev и други 2011) као и паралелизацији овог текста са енглеским оригиналом.

Поред ових пројеката, посебно подршку развоју обраде српског је пружила париска лабораторија LADL (Laboratoire d'Automatique Documentaire et Linguistique) коју је водио проф. Морис Грос⁸. Из ове сарадње је потекла пуна формализација флективне морфологије српског језика и започет је развој морфолошког електронског речника српског на истим принципима који су примењени у изради оваквих речника за француски, енглески и бројне друге језике⁹ (Krstev 2008). Теоријску основу ових речника представља теорија коначних аутомата и трансдуктора (Gross, 1989). У оквиру ове сарадње, београдска група за језичке технологије је овладала програмским алатом за обраду корпуса методом лексичког препознавања Intex (данас Unitex/GramLab)¹⁰ (Слика 4). Овај систем гради интерну репрезентацију текста у облику коначног аутомата, а претрага се врши комплексним упитима који се формулишу на основу садржаја електронских речника у облику коначних трансдуктора, чиме је омогућена не само претрага, већ и модификација корпуса. Из исте школе је преузет и систем CorpusWeb, саставни део програма GlossaNet за надгледање веба (Fairon, 1998), а уз помоћ овог система, био је прикупљен значајан број текстова са раног српског веба, а затим и обрађен помоћу е-речника. У овом окружењу су настали и први већи електронски корпуси српског језика који су обрађивани и експлоатисани локално.



Слика 4. Интерфејс алата Unitex/GramLab са конкорданцама и речником текста

⁸ https://fr.wikipedia.org/wiki/Maurice_Gross

⁹ <https://infolingua.univ-mlv.fr/MenuPrincipal.html#>

¹⁰ <https://unitexgramlab.org/>

3. Први корпус српског језика на вебу

Са продором персоналних рачунара и развојем веба, и међу српским лингвистичким круговима се почело развијати интересовање за поступке аутоматске обраде српског језика. Резултат ове нове атмосфере је, поред осталог било и формирање научно-истраживачког пројекта „Интеракција између текста и речника“¹¹. Један од циљева пројекта је био изградња и постављање на веб корпуса савременог српског језика, пре свега захваљујући настојањима и подршци проф. Љубомира Поповића и проф. Љилане Суботић. Како је пројекат био теоријске природе, „практична“ решења нису била прихватана као резултат рада на пројекту тако да је ова прва верзија корпуса узгредни резултат овог пројекта без одговарајуће валоризације. Са скромним људским и технолошким ресурсима, током 2003. године корпус је компилиран под системом IMS CWB¹² (Utvić, стр. 249) са веб-интерфејсом састављеним на Математичком факултету (Слика 5). За приступ корпусу било је неопходно да се добије корисничка идентификација. Током 20 година постојања, за приступ овој верзији корпуса¹³ је регистровано око 900 корисника из земље и света који су поставили више од милион упита.

Ова верзија данас носи ознаку СрпКор-2 или СрпКор2003. Његова основна намена је била да пружи истраживачима савременог српског језика експерименталну грађу за текућа истраживања, али и да развије културу употребе електронских језичких ресурса. У том светлу, претраживање је у првој верзији корпуса било ограничено на проширене регуларне изразе без других могућности које пружа систем IMS CWB.

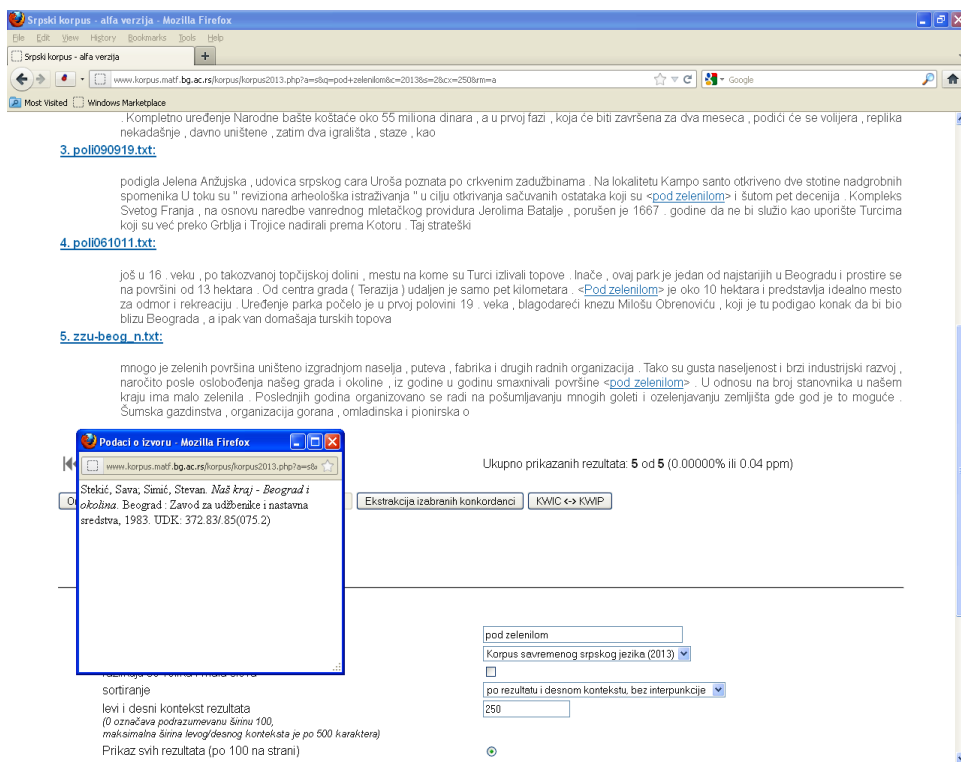
Корпус је садржавао око 23 милиона речи сировог текста без икаквих анотација. Грађа која је унета у корпус, поред различитих извора преузетих са веба, садржала је и материјал који је прикупљан од раних осамдесетих година 20. века, као и изван број литерарних текстова које су донирали њихови аутори. Међу овим текстовима посебно место је чинила преводна књижевност. У корпус је била укључена и уџбеничка литература и мањи бој научних монографија из различитих области. Покушај балансирања садржаја корпуса је морао

¹¹ Пројекат 101743 „Интеракција између текста и речника“ је финансирало Министарство науке и технолошког развоја у периоду 2002-2006. година као пројекат основних истраживања у области језика, а реализатори су били Математички и Филолошки факултети Универзитета у Београду и Филозофски факултет Универзитета у Новом Саду.

¹² <https://cwb.sourceforge.io/index.php>

¹³ Ова верзија је и даље доступна на адреси <http://www.korpus.matf.bg.ac.rs/korpus/login.php> (приступљено 22.04.2025.)

бити напуштен пред дилемом да ли саставити балансиран или велики корпус. Наиме, једномесечна продукција једних дневних новина, нпр. „Политике“, је премашивала у броју речи укупну годишњу домаћу литерарну продукцију.



Слика 5. Веб корпус на Математичком факултету

Проблеми који су се појавили приликом састављања корпуса су обухватили, пре свега, проблем конверзије на јединствену кодну шему ћириличних и латиничних текстова који су могли бити и у различитим кодним шемама. Значајан део грађе, прикупљене са веба, већ следеће године је био постао недоступан преласком сајтова са ISO-889х-серије на Unicode. Такође, садржај корпуса је ограничен на српски језик екавског изговора. За грађу на српском језику ијекавског изговора је планирано да се формира одвојен корпус када се за то стекну услови¹⁴.

¹⁴ Као почетни корак ка формирању корпуса српског језика ијекавског изговора започет је пројекат дигитализације описан у (Шућур & Марковић, 2025).

Даље модификације корпуса су донеле верзију са метаподацима о изворима и пуни репертоар функција у претраживању које пружа систем IMS CWB.

Наставак рада на корпусима савременог српског језика је 2013. донео нову верзију овог корпуса која носи ознаку СрпКор3 или СрпКор2013. Ова верзија, иницијално доступна кроз исти интерфејс као и СрпКор2, је његов прави надскуп проширен текстовима прикупљеним после 2003. године. Овај корпус, који садржи 122 милиона речи, је у технолошком смислу пратио она решења која су била примењена у изградњи СрпКор3, али је значајно обogaћен како у погледу разноврсности текстова, тако и расположивих функција. Овај корпус је лематизиран са обележеном врстом речи и основним метаподацима о изворима. Његов детаљан опис и структура су дати у (Утвић 2013).

4. Даљи развој корпуса и верзије СрпКор-а

Даљи развој пројекта прикупљања текстова и обраде корпуса био је у великој мери вођен еволуцијом корисничких захтева и све израженијим информатичким потребама циљних група корисника корпуса – информатичар с једне стране, а са друге лингвиста, истраживача друштвених наука, педагога и шире јавности заинтересоване за анализу језика. Иницијални захтеви за једноставну претрагу по речима и фразама постепено су прерасли у потребу за комплекснијим упитима који укључују морфолошке, синтаксичке и семантичке критеријуме, као и за напредним статистичким алатима за анализу података из корпуса (нпр. извлачење колокација, стварање фреквенцијских листа, анализа дистрибуције по регистрима, ауторима и сл.).

Паралелно са овим, развој информатичких технологија и појава нових софтверских решења за рад са великим текстуалним подацима учинили су доступним моћније платформе за управљање корпусима. Узимајући у обзир како растуће корисничке захтеве, тако и потребе за ефикасним управљањем значајном количином аотираног текста, донета је одлука о миграцији или имплементацији на платформу која може да пружи напредније функционалности у односу на ранија решења.

Након анализе доступних опција, изабрана је платформа NoSketchEngine (Kilgarriff и други 2014). Ова платформа је препозната као робусно и флексибилно решење које пружа широк спектар алата за корпусну лингвистику, укључујући подршку за комплексне упите користећи језик CQL (Corpus Query Language), могућност креирања

различитих vrsta indeksa i anotacija, kao i napredne alate za analizu. Implementacijom korpusa na platformi NoSketchEngine, omogućeno je ne samo ispušanje trenutnih, već i predviđanje budućih korisničkih zahteva, značajno unapređujući mogućnosti istraživanja i analize dostupnih tekstualnih podataka.

Sketch Engine (Kilgarriff et al., 2004) je softver za korpusnu lingvistiku koji obrađuje unete korpusе и граматичке шаблоне, генеришући детаљне информације о употреби речи (“word sketches”), тезаурусе и поређења блиских синонима (“sketch differences”). То је комерцијални алат који захтева претплату, док NoSketch Engine представља верзију отвореног кода Sketch Engine-а, али са значајним ограничењима у функционалности. За разлику од комерцијалне верзије, NoSketch Engine не долази са готовим корпусима – корисник мора сам да припреми своје корпусе и поседује техничко знање за њихову обраду и форматирање. Недостају многи напредни алати за анализу (као што су “word sketches”, тезаурус, n-грами, трендови), као и алати за аутоматску изградњу корпуса (токенизација, лематизација, тагирање) и графички интерфејс за управљање корпусима. Корисник је у потпуности одговоран за преузимање, инсталацију, хостовање и администрацију NoSketch Engine-а. Подршка је доступна преко Google групе, а постоји и демо инсталација за пробу функционалности.

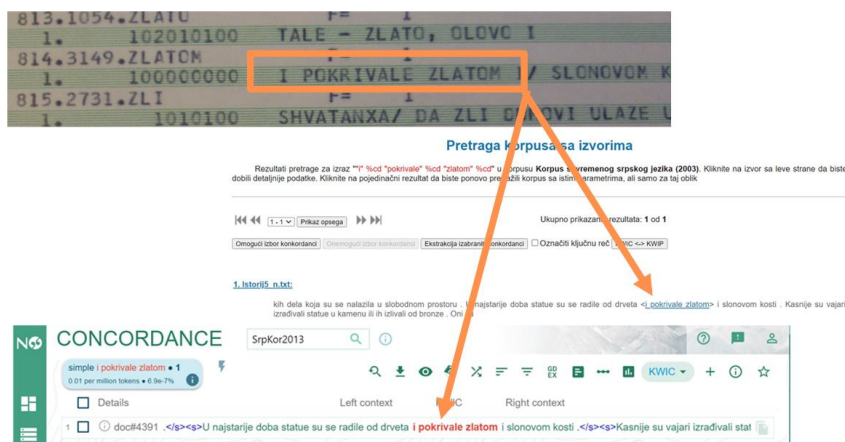
Одржавање и развој корпуса је од 2018. године преузело Друштва за језичке ресурсе и технологије – ЈеРТех¹⁵, које је урадило миграцију корпуса на нову платформу¹⁶, а потом наставило са даљим унапређењима и проширењима текстуалних ресурса. Када је реч о доступним корпусима, са циљем да се у одређеној мери балансирају критеријуми представљања општег језика, доступна су два корпуса. Први од њих је корпус савременог српског језика СрпКор2013 који је преузет са сајта Математичког факултета Универзитета у Београду¹⁷ (поменут у претходном одељку). Током миграције се прешло на латиницу (UNICODE) тако да је замењено ASCII кодирање осмишљено у првом систему за обраду текста АУРОРА. Крајем 2021. године објављена је нови корпус који је допуна претходном и не садржи текстове корпуса СрпКор2013. У поређењу са претходном верзијом (СрпКор2013), СрпКор2021 обухвата више од 600 милиона речи, што га чини једним од највећих корпуса савременог српског језика који има донекле балансиран садржај. Корпус је комплетно аотиран, где је сваки токен обележен припадајућом врстом речи и

¹⁵ <http://jerteh.rs>

¹⁶ <https://noske.jerteh.rs/>

¹⁷ <http://www.korpus.matf.bg.ac.rs/korpus/login.php>

лемом (основним обликом) (Stanković и други 2020). Увођење новог, робуснијег система за управљање корпусним текстовима омогућава његово перманентно унапређивање кроз лако додавање нових корпуса и ажурирање постојећих података. Треба напоменути да се током свих ових миграција водило рачуна да не дође до губитка информација, чему сведочи слика 6, где видимо примере исте конкорданце генерисане системом АУРОРА из прошлог века, а затим у системима СрпКор2 (под IMS CQP) и СрпКор3 (NoSketch).



Слика 6. Животни пут текста

Разноврсност садржаја представља кључну карактеристику фамилије корпуса СрпКор. Поред великог броја новинских текстова, корпус укључује текстове са Википедије, књижевне текстове, уџбенике из широког спектра научних и образовних области, докторске дисертације, као и друге врсте текстова из различитих домена. Ова разноликост обезбеђује бољу репрезентативност различитих регистара и стилова савременог српског језика. У СрпКор унети су и текстови на српском језику из паралелизованих корпуса који су настајали упоредо са СрпКор-ом¹⁸. На овај начин је делимично компензован утицај веб-садржаја на састав корпуса. С друге стране, овакви текстови који су, по правилу, изузетно значајни у културном смислу, јер не само да нису присутни у грађи са веба, већ обично не улазе ни у традиционалне лексикографске корпусе. Њих чине одабрани преводи на српски језик научних, књижевних, филозофских, антрополошких, историјских и сличних текстови преузети из едиција угледних издавачких кућа.

¹⁸ Посебно богата колекција књижевних текстова је преузета из италијанко-српског корпуса.

Корпуси су доступни за претрагу регистрованим корисницима платформе NoSketch на сајту Друштва ЈеРТех¹⁹. Поред СрпКор2021, NoSketch платформа на сајту ЈеРТех подржава и домаће доменске корпусе из специфичних области (математика, рударство, геологија, итд.) демонстрирајући флексибилност и моћ платформе за разноврсна истраживања. Табела 1 приказује квантитативне карактеристике изабраних корпуса. У табели су зеленом бојом засенчени корпуси који се доступни уз бесплатну регистрацију (текстови са Википедије, СрпELTeC потколекција (Krstev и други 2021), паралелни Српско-италијански корпус (Moderc и други 2023)) чиме се подстиче шира употреба платформе, док је за коришћење корпуса СрпКор и осталих корпуса потребна симболична претплата у виду чланарине у Друштву.

Табела 1. Статистика изабраних корпуса на платформи *noske.jerteh.rs*

Корпуси	Токени	Речи	Документи	Реченице
BiKes_sr	1.848.802	1.547.103	336	82.459
BiKes_en	2.072.074	1.766.421	336	82.459
GeoSrpKor	1.316.646	1.067.583	69	
ItSrNER_sr	314.619	254.828	4	10000
ItSrNER_it	331.375	272.026	4	10000
Matematika	1.932.271	1.341.242	89	
RudKorp	3.542.016	2.713.594	172	
SerbItaCor3_sr	12.306.854	10.225.632	267	
SerbItaCor3_it	13.166.565	11.051.704	267	
SkolKor	4.736.884	3.650.849	82	
SrpELTeC	5.923.024	4.766.056	108	144.387
SrFudKo	9.948.301	8.318.399	37	487
SrpKor2013	145.275.324	121.142.927	5.038	5.178.311
SrpKor2021	716.878.652	606.683.682	1.916	29.718.407
WikiKorpus	81.267.249	64.510.812	609.380	609.380
ZlatniKorpus	1.333.804	1.104.864	8	
Укупно	1.002.194.460	840.417.722	618.113	35.835.890

Међу књижевним текстовима посебно се издваја српска потколекција Европске колекције књижевних текстова (ELTeC), која обухвата у основној верзији 100 романа, припремљених у оквиру међународне COST акције CA16204 *Distant Reading for European Literary*

¹⁹ <https://noske.jerteh.rs>

*History*²⁰, чиме се омогућава компаративна анализа са корпусима других европских језика. „Плус“, односно проширена, верзија SrpELTeC+ представља проширење по различитим димензијама, величини, периоду, варијететима текста. Актуелна верзија SrpELTeC+ тренутно садржи више од 190 дела, међу којима су осим романа, путописи и приповетке, а на корпусу се интензивно и даље ради. Важан ресурс свакако представљају резултати прве фазе пројекта “Дигитализација српског књижевног наслеђа ијекавског изговора (1840–1920)” који реализује Центар за дигиталну хуманистику Филозофског факултета Пале, а подржава Министарство за научнотехнолошки развој и високо образовање Републике Српске. Изградња Корпуса српског књижевног наслеђа ијекавског изговора обухватила је 46 публикација, и то приповетке, путописе и романе (Шућур & Марковић 2025).

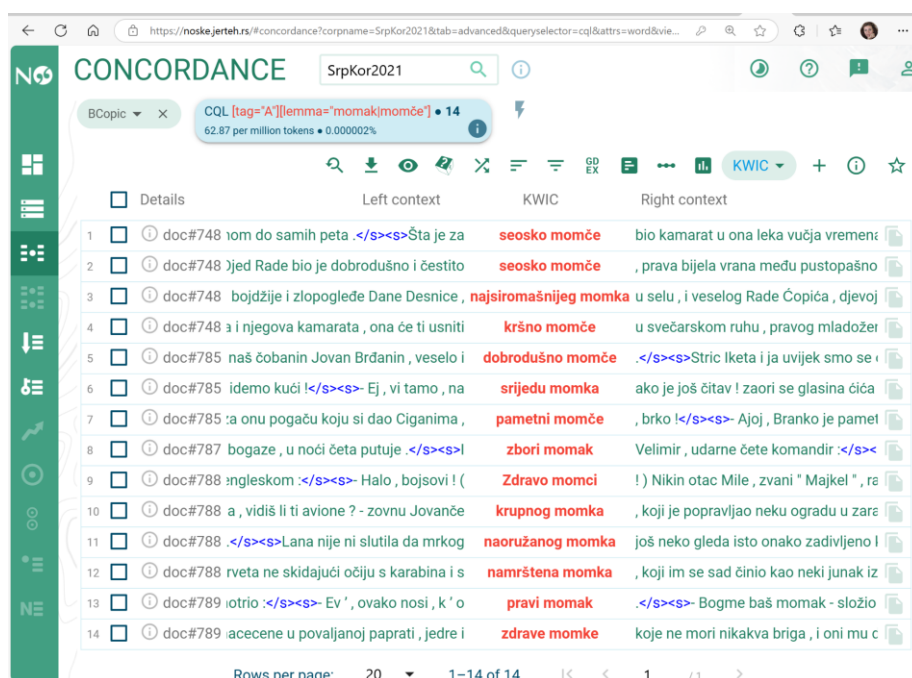
Ниво метаподатака придружених корпусима се разликује, али оно што је заједничко је да су забележени подаци о аутору, години издања, наслову, и регистру текста. Дистрибуција текстова по регистрима у корпусима српског језика из табеле 1 (обједињено), рачунато према броју речи, показује да административни текстови чине 1,3%, књижевност 8,7%, наука 3,2%, новински текстови 67,6%, разговорни 10,4%, уџбеници 0,5%, вики 7,3%, и друго (некласификовано) 1,1%.

На основу метаподатака могуће је креирати подкорпусе и над њима постављати упите, тако да можемо истраживати рецимо само административни језик или, на пример, само језик једног писца. Слика 7 приказује упит на покорпусу VCorpс (издвојена дела Бранка Ћопића) корпуса SrpKor2021 којим се тражи одговор на питање „Који придеви прате именицу *момак*?“, што преведено на језик SQL гласи: `[tag="A"][lemma="momak|momče"]`. Резултат је приказан у облику конкорданци из којих корисник може да види из ког Ћопићевог дела долазе примери и њихов контекст.

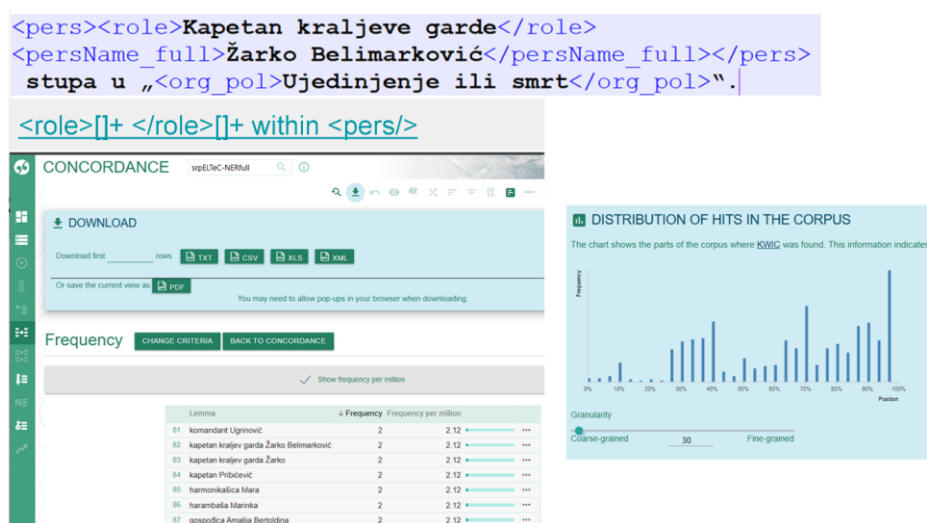
Осим анотација на нивоу речи (лема и врста речи) које су биле поменуте, одређени скуп текстова има и структурне анотације – обележени су одељци, пасуси и реченице, а у неким текстовима и именовани ентитети. Највећи део текстова је обележен са 7 категорија именованих етикета (Stanković и други 2021) одабраних за SrpELTeC корпус (Krstev и други 2021), али један део текстова је обележен и много богатијим скупом етикета коришћењем система за препознавање именованих етикета заснованог на морфолошким речницима и аутоматима (Krstev и други 2019). Слика 8 приказује пример коришћења структурног обележавања у упитима. Упит тражи

²⁰ <https://www.distant-reading.net/>

појављивање етикете <role> (занимања, титуле, позиције) унутар етикете <pers> (особа).



Слика 7. Пример упита на подкорпусу



Слика 8. Пример структурно обележеног текста и упита који укључује структурне етикете

5. Интеракције СрпКор-а са другим ресурсима

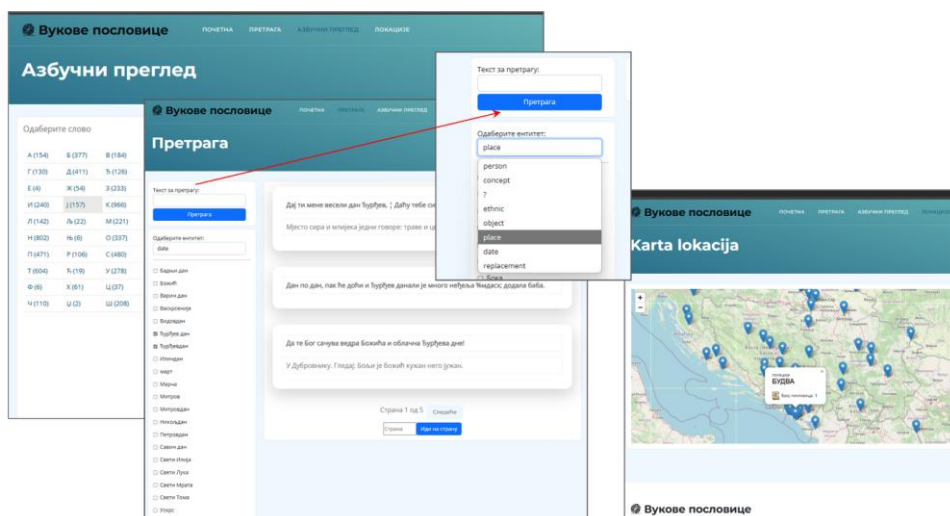
Интеракција корпуса савременог српског језика (СрпКор) са другим језичким ресурсима представља кључни аспект модерне лингвистике и обраде природних језика. Ова спрега, првенствено са речницима који пружају структурирано лексичко знање и језичким моделима који обухватају дистрибутивне и семантичке односе, омогућава обострано обогаћивање. Корпус пружа емпиријску основу за валидацију и проширење речника, док речници и модели унапређују могућности корпусних анализа и анотације.

Успостављена је снажна двосмерна спрега између лексичке базе Лексмирка и фамилије корпуса СрпКор (Lazić, Škorić, 2020). Напоменимо да је база Лексмирка настала трансформацијом електронских речника за српски језик (Krstev 2008) у релациону базу (Stanković и други 2018) прилагођену за даље серијализације у екосистем повезаних лингвистичких података. Ова интеракција омогућава обострано обогаћивање ресурса. Са једне стране, кроз специјализовани интерфејс лексичке базе Лексмирка, корисници имају директан и интерактиван увид у примере употребе речи и њихових различитих облика у аутентичном корпусном контексту (слика 9). Такође је могуће истраживати појављивање речи у специфичним синтаксичким обрасцима или колокацијама детектованим на корпусу. Са друге стране, лексичка база и придружени морфолошки ресурси имају кључну улогу у унапређењу система за лематизацију корпуса. Систем за лематизацију се континуирано унапређује из верзије у верзију корпуса, а његова прецизност и обухватност директно зависе од квалитета и обухвата електронских морфолошких речника српског језика. Ови речници, коришћењем моћног система за обраду текста, попут Unitex-а, омогућавају не само похрањивање лема и њихових основних морфолошких информација, већ и прецизно генерисање свих валидних флективних облика сваке леме. Ова свеобухватна морфолошка информација је затим интегрисана у процес анотације корпуса, обезбеђујући да се свака појава речи у тексту исправно повеже са својом основном лемом.

Речнички чланци базе Лексмирка повезани су и са Ворднетом (Krstev и други 2004, Stanković и други 2018), терминолошком базом Терми (Kitanović и други 2021), одређеним бројем дигитализованих традиционалних речника и речницима са веба.

Слика 9. Интерфејс лексикона и корпуса на примеру Лексимирке и СрпКор-а: проналажење разноврсних информација полазећи од упита “јужнословенски” у систему Лексимирка.

Напоменимо још један пример текста који је прошао вишедеценијску трансформацију и добио ново „рухо“. Дигитализација Вукових пословица и израда њиховог индекса део су резултата докторске дисертације “Један прилаз информатичком моделирању текста и алгоритми његове трансформације” (1997) проф. др Цветане Крстев (Krstev 1996, Krstev 1997). Из изворног SGML формата који је пратио смернице TEI, генерисан је XML облик на основу ког су подаци преточени у PostgreSQL релациону базу. Развијена је веб апликација НарПос²¹ (слика 10), уз помоћ које се могу прегледати пословице уређене у азбучном поретку, а могу се и, претраживати преко индекса, филтрирати по темама али и проналазити на карти, на основу информације која се налази у самом тексту пословице или у објашњењу које је прати).



Слика 10. Веб апликација Народне пословице

Развој савремених језичких модела, укључујући оне за обраду и разумевање српског језика, фундаментално зависи од доступности и квалитета великих текстуалних корпуса. Иако веб-корпуси нуде неупоредиву количину података и ефикасно одражавају тренутну, често неформалну, употребу језика на интернету, њихова нехомогеност, присуство грешака, сленга, нестандардних варијанти и недостатак конзистентне, дубинске лингвистичке анотације могу представљати ограничење за тренирање модела који захтевају разумевање структурираног и формалнијег језика. Због тога су

²¹ <https://narpos.jerteh.rs/>

пажљиво прикупљени, балансирани и анотирани корпуси, чији садржај не потиче искључиво са веба (већ укључује и књижевна дела, научне радове, званичне документе, транскрибовани говор, итд.), незаменљиви за изградњу робусних језичких модела способних да обрађују различите регистре и стилове. Ови “чистији” и структуриранији корпуси пружају поузданију основу за учење сложених граматичких образаца, нијанси семантике и специфичности различитих домена, што је од пресудног значаја за развој модела високе прецизности и поузданости, посебно за задатке који превазилазе површинску обраду текста. Подразумева се да је припрема оваквих корпуса далеко захтевнија како у фази прикупљања текстова тако и у фази њихове обраде која претходи укључивању у корпус.

Пројекат ТЕСЛА²² посебно се бави векторизацијом текста – представљањем речи, фраза или читавих докумената као нумеричких вектора, што је темељна техника у савременој обради природних језика. Овим се отварају врата за развој низа напредних апликација за српски језик – од анализе осећања и ставова и класификације текста до машинског превођења и система за одговарање на питања. Успешност ових апликација директно корелира са доступношћу и квалитетом великих колекција текстуалних података, а посебно анотираних корпуса који пружају лингвистички обogaћене информације. У том контексту, планови за будући развој корпуса СрпКор усмерени су ка стварању још већег и детаљније анотираног ресурса.

6. Закључак

У овом раду пружамо свеобухватан преглед развојног пута корпуса савременог српског језика. Полазећи од почетних фаза конципирања и прикупљања података, пратимо стварање првог корпуса српског језика доступног путем веб-технологија, што је представљало значајан корак у дигитализацији и доступности језичких ресурса за српски. Детаљно елаборирамо даљи развој и различите итерације корпуса, које су донеле значајна проширења у обиму, разноврсности садржаја и дубини лингвистичке анотације, одражавајући еволуцију истраживачких потреба и технолошких могућности. Коначно, представљамо актуелну фазу развоја, фокусирану на интеграцију и синергију корпуса са другим значајним језичким ресурсима, као што су лексичке базе и језички модели, што

²² <https://tesla.rgf.bg.ac.rs/>

омогућава обострано обогаћивање и отвара нове могућности за напредне језичке анализе и апликације. Овај преглед документује трансформацију корпуса од пионирског подухвата до савременог, интероперабилног ресурса, илуструјући изазове и решења током више декада развоја.

Даљи развој корпуса савременог српског језика, усмерен ка стварању робустне и детаљне лингвистичке основе за напредне апликације попут векторизације текста, планиран је кроз више паралелних активности. Крајем 2025. године планирано је објављивање нове верзије корпуса на платформи NoSketch, која ће обухватити значајно проширење садржаја. Ово проширење укључиће новинске текстове из периода 2022-2025, као и нове типове текстова попут путописа, мемоара и историјских новина, проширујући временску димензију и доприносећи културно-историјској вредности корпуса, уз континуирано допуњавање одабраним веб-садржајима. Паралелно са повећањем обима, фокус је стављен на детаљније обогаћивање метаподатака. Ово ће омогућити флексибилније креирање подкорпуса према различитим димензијама, укључујући период, домен, изговор, ауторе или специфичне типове текста, чиме ће се омогућити прецизнија циљана истраживања. Квалитет корпуса ће бити додатно унапређен кроз допуну и обогаћивање лингвистичких анотација. У плану је комплетирање поделе на реченице, препознавање именованих ентитета, те додавање детаљнијих граматичких информација и повезивање са базама знања. У том циљу, активно се ради на развоју и унапређењу модела и алата за аутоматску анотацију који ће подржати, не само стандардно означавање врста речи (PoS) и лематизацију, већ и сложеније задатке попут препознавања именованих ентитета, додељивања значења речима и креирања веза ка базама знања. Сви ови напори усмерени су ка изградњи корпуса који ће пружити незаменљив ресурс за лингвистичка истраживања и развој напредних алата за обраду српског језика.

Захвалница: Ово истраживање је подржао Фонд за науку Републике Србије, #7276, Text Embeddings - Serbian Language Applications - TESLA.

Литература

- [1] Fairon, C. (1998). GlossaNet: Parsing a web site as a corpus. *Linguisticæ Investigationes*, 22(1-2), 327–340.
- [2] Gross, M. (1989). The Use of Finite Automata in the Lexical Representation of Natural Language; in: Gross, M., Perrin, D. (Eds.). *Electronic Dictionaries and Automata in Computational Linguistics*, Lecture Notes in Computer Science. Berlin: Springer Verlag.
- [3] Dobrić, N. (2012) "Savremeni jezički korpusi na Zapadnom Balkanu–Istorijat, trenutno stanje i budućnost." *Slavistična revija* 60: 677-692.
- [4] Ivić, M. (2001). *Pravci u lingvistici* (Vol. 2, 9. izd.). Čigoja štampa, Biblioteka XX vek, Beograd.
- [5] Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., ... & Suchomel, V. (2014). The sketch engine. *Lexicography*, 1(1), 7-36.
- [6] Kilgarriff, A., Rychlý, P., Smrž, P., & Tugwell, D. (2004). The Sketch Engine. *Proceedings of the 11th EURALEX International Congress* (crp. 105-116).
- [7] Kitanović, O., Stanković, R., Tomašević, A., Škorić, M., Babić, I., & Kolonja, L. (2021). A data-driven approach for raw material terminology. *Applied Sciences*, 11(7), 2892.
- [8] Krstev, C. (1996) [računarska obrada leksike], *Vukove narodne poslovice s registrom ključnih reči*, ed. Slobodan Đorđević, Nolit, Beograd, 1996.
- [9] Krstev, C. (1997). *Jedan prilaz informatičkom modeliranju teksta i algoritmi njegove transformacije* (doktorska disertacija, Univerzitet u Beogradu, Matematički fakultet).
- [10] Krstev, Cvetana. *Processing of Serbian. Automata, texts, and electronic dictionaries*. Faculty of Philology of the University of Belgrade, 2008.
- [11] Krstev, C. Pavlović-Lažetić, G., Vitas, D., & Obradović, I. (2004) Using Textual and Lexical Resources in Developing Serbian Wordnet, *Romanian Journal of Information Science and Technology*, Romanian Academy, Publishing House of the Romanian Academy, 7(1-2), 147-161.
- [12] Krstev, C., Obradović, I., Utvić, M., & Vitas, D. (2014). A system for named entity recognition based on local grammars. *Journal of Logic and Computation*, 24(2), 473-489.
- [13] Krstev, C., Vitas, D., & Erjavec, T. (2004). MULTTEXT-East Resources for Serbian. V T. Erjavec & J. Zganec Gros (Urednici), *Zbornik 7. međunarodne multikonferencije "Informacijska družba IS 2004", Jezikovne tehnologije*. Institut "Jožef Stefan".
- [14] Krstev, C., & Vitas, D. (2005) "Corpus and Lexicon - Mutual Incompleteness", *Proceedings of the Corpus Linguistics Conference*, 14-17 July 2005,

- Birmingham, eds. Pernilla Danielsson and Martijn Wagenmakers, ISSN 1747-9398, <http://www.corpus.bham.ac.uk/PCLC/>.
- [15] Krstev, C., Vitas, D., & Trtovac, A. (2011). Orwell's 1984 – the Case of Serbian Revisited. Y Z. Vetulani (Yp.), *Proceedings of 5th Language & Technology Conference* (стр. 570–574). Fundacija Uniwersytetu im. A. Mickiewicza.
 - [16] Krstev, C., & Stanković, R. (2022). *Deliverable D1.35 Report on the Serbian Language. European Language Equality (ELE)*; EU project no. LC-01641480–101018166. <https://european-language-equality.eu/reports/language-report-serbian.pdf>.
 - [17] Krstev, C., & Stanković, R. (2023). Language Report Serbian. In: Rehm, G., Way, A. (eds) *European Language Equality*. Cognitive Technologies. Springer, Cham. https://doi.org/10.1007/978-3-031-28819-7_32
 - [18] Lazić, B., & Škorić, M. (2020). From DELA based dictionary to Leximirka lexical database. *Infotheca - Journal For Digital Humanities*, 19(2), 81-98. doi:10.18485/infotheca.2019.19.2.4
 - [19] Moderc, S., Stanković, R., Tomašević, A., & Škorić, M. (2023). An italian-serbian sentence aligned parallel literary corpus. *Review of the National Center for Digitization*, 43 (2023), 78–91.
 - [20] Pierce, J. R., & Carroll, J. B. (1966). *Language and machines—Computers in translation and linguistics* [ALPAC report]. National Academy of Sciences, National Research Council, Washington, DC.
 - [21] Rajić, Lj. (1981). *Teorija i poetika prevođenja*. Prosveta, Beograd.
 - [22] Stanković, R., Krstev, C., Todorović, B. Š., & Škorić, M. (2021). Annotation of the serbian eltec collection. *Infotheca—Journal for Digital Humanities*, 21(2), 43-59.
 - [23] Stanković, R., Krstev, C., Lazić, B., & Škorić, M. (2018). Electronic Dictionaries—from File System to Lemon-based Lexical Database. In the *6th Workshop on Linked Data in Linguistic (LDL-2018)*, Towards Linguistic Data Science.
 - [24] Stanković, R., Mladenović, M., Obradović, I., Vitas, M., & Krstev, C. (2018). Resource-based WordNet augmentation and enrichment. Y *Proceedings of the Third International Conference on Computational Linguistics in Bulgaria (CLIB 2018)* (стр. 104-114).
 - [25] Stanković, R., Sandrih, B., Krstev, C., Utvić, M., & Škorić, M. (2020). Machine learning and deep neural network-based lemmatization and morphosyntactic tagging for serbian. In LREC 2020-12th International Conference on Language Resources and Evaluation, Conference Proceedings (pp. 3954-3962). European Language Resources Association (ELRA).

- [26] Šipka, M. (Ур.). (1978). *Kompjuterska obrada lingvističkih podataka*, Posebna izdanja 4. Institut za književnost i jezik, Sarajevo.
- [27] Tomić, T. (1978). Statistička analiza srpskohrvatskog teksta pomoću računara. У М. Šipka (Ур.), *Kompjuterska obrada lingvističkih podataka* (стр. 221–236). Institut za književnost i jezik.
- [28] Utvić, M. (2013). *Izgradnja referentnog korpusa savremenog srpskog jezika*. Doktorska disertacija. Univerzitet u Beogradu, Filološki fakultet.
<https://nardus.mpin.gov.rs/handle/123456789/4091>
- [29] Vitas, D. & Krstev C. (2012) “Tvorbeni obrasci u elektronskom rečniku srpskog jezika”, Međunarodni komitet slavista. Komisija za tvorbu reči. *Međunarodna naučna konferencija Tvorba reči i njeni resursi u slovenskim jezicima* (14), pp. 515-525, Filološki fakultet Univerziteta u Beogradu, Beograd, ISBN 978-86-6153-116-3
- [30] Vitas, D. (1979) Prikaz jednog sistema za automatsku obradu teksta, *INFORMATICA'79*, Bled, (стр. 7 10)
- [31] Vitas, D. (1980) Generisanje imeničkih oblika u srpskohrvatskom, *Informatica* 80(3) Slovenačko društvo za informatiku, Ljubljana, 49-55.
- [32] Витас, Д. (2023) Белешке о ручној и аутоматској обради српског језика, *Језик Данас*, бр. 22, 2023, Матица Српска, Нови Сад.
- [33] Витас, Д. (2018) О развоју информатике међу математичарима, *Инфотека* 18 (1), 2018,
- [34] Витас, Д., Поповић, Ј. (2003) „Конспект за изградњу референтног корпуса српског стандардног језика“, *Научни састанак слависта у Вукове дане 31/1 - МСЦ*, Београд, стр. 221 - 227.
- [35] Шућур, С., Марковић, Ј. (2025). Дигитализација српског књижевног наслеђа ијекавског изговора (1840–1920) при Центру за дигиталну хуманистику Филозофског факултета Пале (прва фаза), Зборник радова конференције Јужнословенски језици у дигиталном окружењу Јудиг 2025 (у овом броју), Универзитет у Београду, Филолошки факултет.

About the Family of the Corpus of the Modern Serbian Language SrpKor

Duško Vitas, Ranka Stanković, Cvetana Krstev

Summary

The SrpKor family comprises electronic corpora of contemporary Serbian, initiated in the late 1970s and first made publicly available online in 2003. Its development, especially before the advent of abundant web-based resources, involved systematic text collection, corpus processing tool creation, and annotation methodology design. SrpKor is not merely a digital text collection but an integrated resource with software infrastructure, linguistic annotations (metadata, morphological tagging, lemmatization, named entities), and search capabilities, evolving through multiple versions under constrained resources.

Early Serbian NLP efforts in the 1980s lagged behind other European contexts, partly due to limited computational literacy among philologists and skepticism toward language technology. Initial advances came from computer scientists, leading to the Aurora corpus processing system and early morphological analyzers. International collaborations in the 1990s (e.g., Telri, LADL) introduced formal morphological modeling and tools such as Unitex/GramLab, facilitating large-scale corpus compilation.

The first web corpus (SrpKor2003) contained ~23 million unannotated words; SrpKor2013 expanded to 122 million words, with lemmatization and part-of-speech tagging. Responding to growing research needs, the corpora migrated to the NoSketchEngine platform, culminating in SrpKor2021—a 600M-word, fully annotated balanced corpus. These corpora include diverse genres: news, literature, Wikipedia, textbooks, academic texts, parallel translations, and culturally significant non-web materials. Metadata enables subcorpus creation by domain, author, or register.

SrpKor interacts closely with lexical databases such as Leximirka, enhancing both lexicon and corpus quality. Structural and named entity annotations support complex queries. Additional domain-specific corpora (e.g., mining, geology, mathematics) demonstrate the platform's versatility.

Future development focuses on expanding temporal coverage (including 2022–2025 texts), enriching metadata, refining linguistic annotation (sentence segmentation, detailed grammar, named entities), and

linking to knowledge bases. These efforts aim to provide a robust, richly annotated resource supporting advanced applications like text vectorization, sentiment analysis, and machine translation, ensuring SrpKor remains a cornerstone for Serbian language research and technology development.

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

Models for Automatic Morphological Inflection of Serbian and Croatian Based on the srLex and hrLex Morphological Lexicons

Scientific paper

DOI: 10.18485/judig.2025.1.ch2

Jaka Čibej¹  0000-0002-3037-6848

Abstract

Open-source machine-readable morphological lexicons are useful for morphosyntactic tagging of corpora and represent a crucial step toward compiling modern digital dictionary databases. In the paper, we present the first step toward extending the functionalities of Pregibalnik, a custom developed open-source tool for Slovene lexicon expansion, to cover Serbian and Croatian and help automatically expand the lexicons with new entries. We describe the process of extraction of morphological patterns from the hrLex and srLex inflectional lexicons of Croatian and Serbian, as well as a robust process of feature selection based on ending word parts. The features are used to develop a series of machine-learning models to predict morphological patterns for Croatian and Serbian lexemes, achieving an average F1-micro score of 0.85 (depending on lexeme type). This also helps identify potential inconsistencies within the current versions of the lexicons. The extracted patterns and models are available under a Creative Commons CC-BY 4.0 license.

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

1. Introduction

Open-source machine-readable morphological lexicons are not only helpful for human users (particularly for those studying highly inflectional languages as a second language) but are a useful resource for a wide range of tasks in natural language processing and computational linguistics. They can be

¹ Centre for Language Resources and Technologies, University of Ljubljana, Faculty of Arts, University of Ljubljana, jaka.cibej@ff.uni-lj.si

used to improve morphosyntactic tagging of corpora and represent a crucial step toward compiling modern digital dictionary databases. An example is the *Digital Dictionary Database of Slovene* (DDDS; Kosem et al., 2021), an open-access lexicographic relational database that is being developed at the Centre for Language Resources and Technologies of the University of Ljubljana. The morphological basis for DDDS is the *Sloleks Morphological Lexicon of Slovene* (Čibej et al., 2022). In the RSDO (Development of Slovene in a Digital Environment)² project, version 2.0 with approximately 100,800 lexemes was updated to version 3.0 by adding approximately 265,000 new lexemes from the *Gigafida 2.0 Corpus of Written Standard Slovene* (Krek et al., 2020), along with their inflected forms, accentuated forms, and IPA/SAMPA pronunciations. All were automatically generated using *Pregibalnik*³ ("Inflector" in English; from the Slovene verb *pregibati* 'to inflect'), a custom-developed open-source tool for Slovene lexicon expansion (more on this in Section 2). Sloleks is also used in the development of the Slovene CLASSLA-Stanza models for lemmatization (Terčon et al., 2023) and morphosyntactic tagging (Ljubešić et al., 2023).

Two open-source lexicons similar to Sloleks have been published for Serbian and Croatian – srLex 1.3 (Ljubešić 2019a) and hrLex 1.3 (Ljubešić 2019b), compiled from srWaC (Ljubešić & Klubička, 2016a) and hrWaC (Ljubešić & Klubička, 2016b) web corpora, respectively. Similar to Sloleks, srLex and hrLex are also used in the Serbian and Croatian CLASSLA-Stanza models, which is why it is important to keep the lexicons up-to-date and extend them with new lexemes. Because Croatian and Serbian are structurally similar to Slovene⁴ and because they share a similar infrastructural framework, the same method applied to Slovene data can be used (with some minor adjustments) to extend the functionalities of *Pregibalnik* to also cover Croatian and Serbian. However, while machine-learning methods for lexicon expansion have already been used to predict paradigms for Croatian and Serbian, the results are either not available under an open-access license (see Šnajder, 2013) or are not directly compatible with the infrastructure of *Pregibalnik*: for instance, the machine-readable paradigms used by Ljubešić et al., 2016 were only available in the

² RSDO Project Site: <https://rsdo.slovenscina.eu/>

³ The code for *Pregibalnik* is available on Github: <https://github.com/clarinsi/SloInflector>

Pregibalnik is also available as an API service:

<https://orodja.cjvt.si/pregibalnik/redoc>

<https://orodja.cjvt.si/pregibalnik/docs>

<https://orodja.cjvt.si/pregibalnik/form-generator/docs>

<https://orodja.cjvt.si/pregibalnik/form-generator/redoc>

⁴ In this paper, we treat Serbian and Croatian as completely separate because we use different resources (srLex and hrLex, respectively) to develop their inflectional models. This is a purely pragmatic decision made in accordance with the infrastructure of *Pregibalnik* and is not intended as a reflection of the linguistic continuum in actual language use.

Apertium format,⁵ which for instance sometimes does not clearly distinguish between morphological patterns for masculine, feminine, and neuter nouns, which according to the MULTEXT-East Morphosyntactic Specifications (MTE)⁶ used by *Pregibalnik* are lexeme-level features that clearly discriminate between morphological patterns.

In this paper, we present the first step toward extending the functionalities of *Pregibalnik* to cover Serbian and Croatian and help automatically expand the lexicons with new lexemes using an easily accessible API service. The paper is structured as follows: we first present the structure of *Pregibalnik* focusing on the form generator component (Section 2), then describe the process of extracting morphological patterns from srLex and hrLex (Section 3) and the features used in predictions (Section 4). We evaluate the developed models (Section 5) and provide a brief qualitative analysis of some of the most frequent misclassifications (Section 6), then conclude the paper with some suggestions for future work (Section 7).

2. Lexicon Expansion with Pregibalnik

Pregibalnik currently consists of three components which can be used separately or as part of a single process: the form generator, the accentuator, and the IPA/SAMPA grapheme-to-phoneme converter. The workflow is shown in Figure 1.

The tool takes a lemma and its MTE lexeme-level morphosyntactic features (e.g. the Slovene word *omikron* 'omicron' noun, common, masculine) as input and first generates a complete paradigm of forms inflected by case, number, tense, etc. (nominative singular *omikron*, genitive singular *omikrona*, dative singular *omikronu*, and so on). This is then forwarded to the other two components to add accentuated forms (*ómikron*) and pronunciations (IPA: /'o:mikrɔn/) as well. In this paper, we focus on form generation for Croatian and Serbian as srLex and hrLex currently only include inflected forms.

The first component of *Pregibalnik* generates the set of forms by first extracting a set of features from the input lemma in the form of a numeric vector. For the Slovene form generation models, the features are mostly based on a linguistically informed list of ending word parts (mostly suffixes used in word formation, e.g. '*acija*' in *liofilizacija* 'lyophilization') as well as several other features, such as the ratio of upper-case and lower-case characters (e.g. to help detect acronyms such as *ZN* (*Združeni narodi* 'United Nations'), which

⁵ Croatian-Bosnian-Serbian paradigms are available at: <https://sourceforge.net/p/apertium/svn/HEAD/tree/languages/apertium-hbs/apertium-hbs.hbs.metadix>

⁶ MULTEXT-East Morphosyntactic Specifications for Slovene: <https://nl.ijs.si/ME/V6/msd/html/msd-sl.html>

are inflected differently compared to other nouns in the same category. The numeric vector is then fed into one of several models (based on the part-of-speech of the relevant lexeme) that predicts the code of the morphological pattern, a blueprint consisting of pairs of MTE morphosyntactic tags and their ending word parts. The pattern is then used to generate the entire paradigm. The workflow is presented in Figure 2.

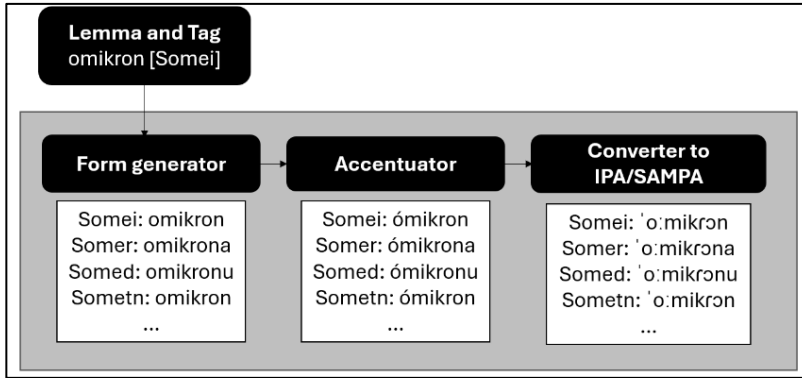


Figure 1: Representation of the *Pregibalnik* workflow for the Slovene masculine common noun *omikron* 'omicron'.

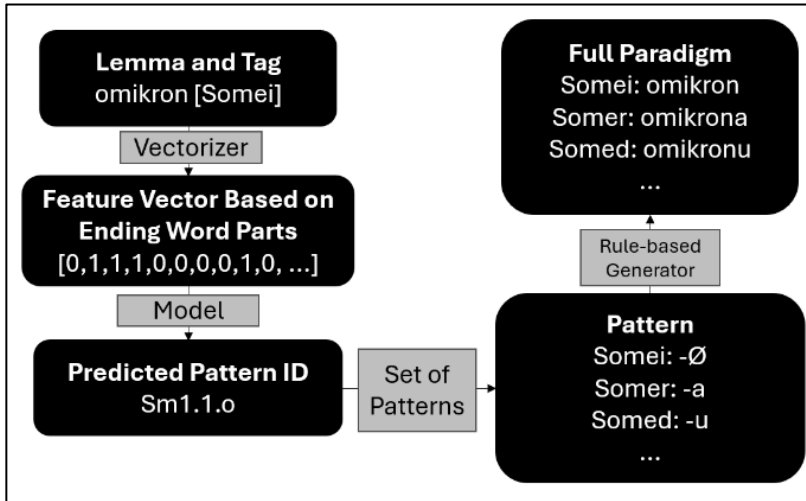


Figure 2: Form generation workflow in *Pregibalnik*.

The set of machine-readable Slovene morphological patterns were automatically extracted from Sloleks using a robust approach (see Section 3 for a more detailed description on the example of hrLex and srLex). The

patterns were then manually validated and hierarchically sorted (Arhar Holdt & Čibej 2018; Arhar Holdt 2021) before being used in machine-learning predictions.

We performed the same bottom-up process of morphological pattern extraction on hrLex and srLex. However, we test a more agnostic approach with no pre-defined list of word ending parts for pattern predictions and no hierarchization, which requires some additional manual work and linguistic expertise.

3. Morphological Pattern Extraction

The hrLex 1.3 and srLex 1.3 lexicons consist of approximately 164,000 and 169,000 lexemes,⁷ respectively, and contain the following data: word forms, their lemmas, morphosyntactic tags and features according to the Serbo-Croatian MULTEXT-East (MTE) Morphosyntactic Specifications⁸, morphosyntactic tags and features according to the Universal Dependencies annotation scheme, and the absolute and relative frequencies of the form-lemma-tag combination from the corpus (hrWaC and srWaC for hrLex and srLex, respectively). An excerpt from hrLex is shown in Table 1.

Table 1: Excerpt from hrLex.

Form	Lemma	MTE Tag	MTE Features	UD Tag	UD Features	f _A	f _R
hljeba	hljeb	Ncmmsg	Type=common Gender=masculine Number=singular Case=genitive	NOUN	Case=Gen Gender=Masc Number=Sing	588	0.000421
hljeb	hljeb	Ncmnsn	Type=common Gender=masculine Number=singular Case=nominative	NOUN	Case=Nom Gender=Masc Number=Sing	269	0.000192
hljebu	hljeb	Ncmstd	Type=common Gender=masculine Number=singular Case=dative	NOUN	Case=Dat Gender=Masc Number=Sing	2	0.000001

The process of extracting morphological patterns was based on a simple algorithm that first searches for all forms pertaining to a lexeme,

⁷ A lexeme is an entry in the lexicon, consisting of the lemma form, its inflectional forms, and their morphosyntactic features.

⁸ MULTEXT-East Morphosyntactic Specifications for Serbo-Croatian (v6):
<https://nl.ijs.si/ME/V6/msd/html/msd-hbs.html>

then first identifies the immutable part, i.e. the part that is common to all the forms (Table 2).

*Table 2: Tags and forms with immutable parts (in bold) for the lexeme *abakus* (noun, common, masculine) from hrLex 1.3.*

Singular Forms	Plural Forms
Ncmsn: abakus	Ncmpn: abakusi
Ncmsg: abakusa	Ncmpg: abakusa abakusâ
Ncmsd: abakusu	Ncmpd: abakusima
Ncmsan: abakus	Ncmpa: abakuse
Ncmsv: abakuse	Ncmpv: abakusi
Ncmssl: abakusu	Ncmpl: abakusima
Ncmssi: abakusom	Ncmpi: abakusima

The immutable part is then removed from the forms to determine the mutable parts for each tag and create the blueprint for the morphological pattern pertaining to the lexeme (Table 3). Each unique morphological pattern is assigned an ID (formatted as P_{lexicon}_{lexeme-level features}_{sequential_number}, e.g. P_hrLex_Ncm_1).

*Table 3: Morphological pattern extracted from the lexeme *abakus* (noun, common, masculine) from hrLex 1.3.*

Singular Forms	Plural Forms
Ncmsn: -Ø	Ncmpn: -i
Ncmsg: -a	Ncmpg: -a -â
Ncmsd: -u	Ncmpd: -ima
Ncmsan: -Ø	Ncmpa: -e
Ncmsv: -e	Ncmpv: -i
Ncmssl: -u	Ncmpl: -ima
Ncmssi: -om	Ncmpi: -ima

The results of the pattern extraction are shown in Table 4. The difference in the number of extracted patterns between hrLex and srLex is immediately apparent, with srLex accounting for three times the number of patterns extracted from hrLex. The reason for this discrepancy lies in the treatment of the combinations of Ekavian and Ijekavian forms in srLex, where both spelling variants are included as part of the same lexeme (e.g. the lexeme *cenovnik* 'price list' contains like *cenovnik* and *cjenovnik*),

which results in a great number (up to 60 %) of single-occurrence patterns (for instance, the immutable part in the lexeme containing both *cenovnik* and *cjenovnik* is *c-*, while the mutable parts are *-enovnik/-jenovnik*, which do not fit any other morphological pattern). This raises the question of the manner of including Ekavian and Ijekavian forms in the lexicons. They should arguably be treated as separate lexemes since Ekavian and Ijekavian phenomena are not part of inflectional morphology, but rather variants of lexemes with the same morphological patterns.

Table 4: Extracted morphological patterns from *hrLex* and *srLex*.

Lexeme type	Patterns in <i>hrLex</i> 1.3	Patterns in <i>srLex</i> 1.3
Noun, common, masculine (Ncm)	284	552
Noun, common, feminine (Ncf)	81	406
Noun, common, neuter (Ncn)	44	272
Noun, proper, masculine (Npm)	178	178
Noun, proper, feminine (Npf)	47	48
Noun, proper, neuter (Npn)	11	11
Verb, main (Vm)	254	466
Adjective, general (Ag)	173	656
Adjective, possessive (As)	3	361
Adjective, participial (Ap)	24	140
Adverb, general (Rg)	136	616
Adverb, participial (Rr)	44	239
Total	1,279	3,945

4. Prediction Features Based on Typical Ending Word Parts

To construct the set of features for predicting morphological patterns, we first export frequency lists of ending word parts (1-5-grams) from both lexicons for each lexeme type (common masculine nouns, general adverbs, etc.). We compare the frequency (f_A) of each ending word part (a) within each morphological pattern (P) to its frequency outside the morphological pattern (f_B) to obtain the pattern typicality score (S), which indicates how typical the ending word part is for pattern P :

$$S(a, P) = \frac{f_A + 0.01}{f_B + 0.01}$$

We then calculate the global typicality score (G) of each ending word part by comparing the maximum and average pattern typicalities across morphological patterns:

$$G(a) = \frac{S_{max} + 0.01}{S_{avg} + 0.01}$$

For each lexeme type, we thus obtain a list of ending word parts along with their absolute frequencies and global typicality scores. The higher the ratio between the maximum and average is, the more typical the ending word part is for a specific morphological pattern, which indicates that the ending word part can contribute toward discriminating between different patterns.

Table 5 shows the top 10 ending word parts for common feminine nouns in hrLex 1.3 sorted by frequency. Ending word parts such as *-a*, *-ca*, and *-ica* are less useful for discriminating between patterns, whereas *-ja*, *-ija*, *-cija* on the one hand and *-t*, *-st*, *-ost*, and *-nost* on the other feature higher typicality scores.

Table 5: Top 10 ending word parts for common feminine nouns in hrLex 1.3.

Ending word part	Absolute frequency	Global Pattern Typicality
a	10,203	43.78
t	2,952	79.88
st	2,924	80.51
ost	2,872	80.57
ja	2,805	76.81
ca	2,483	46.83
ica	2,364	50.03
ija	2,323	79.50
nost	2,104	80.53
cija	1,182	80.83

We made a selection of ending word parts for each of the 12 lexeme categories. We removed ending word parts that occur in less than 10 lemmas and kept the first 500 word parts sorted by typicality (or all of the relevant ending word parts if the list contained less than 500 word parts). We compiled two separate vectorizers (one for each language) that use the lists of ending word parts from the relevant lexicon (hrLex or srLex) to construct a numeric vector from the input lemma. The vector of each

relevant lexeme is then paired with the morphological pattern code to compile the training data for machine learning models, which we present in more detail in the following section.

5. Model Training and Quantitative Evaluation

We trained separate models for each lexeme type to avoid any unnecessary misclassification errors on the level of parts-of-speech – a single model trained on all morphological patterns regardless of their lexeme-level features could potentially assign e.g. an adverbial pattern to a verb or vice-versa.

Four model architectures⁹ were considered, as shown in Table 6. For each model type and each language, 12 models were trained for each lexeme type, and evaluation scores were aggregated across different patterns. which lists F1-micro scores over all morphological patterns. We list F1-micro scores here to present the overall model performance on the lexicon, not an average across different morphological patterns as some classes are very infrequent and are likely the results of errors in the lexicon rather than linguistic idiosyncrasies that need to be accurately predicted.

Table 6: F1-micro scores for morphological pattern classification in hrLex 1.3 and srLex 1.3.

Model	F1-micro (hrLex 1.3)	F1-micro (srLex 1.3)
k Neighbors Classifier (k=5)	0.8366	0.8317
Linear Support Vector Classifier	0.8534	0.8553
Logistic Regression	0.8607	0.8507
Multinomial Naïve Bayes Classifier	0.8467	0.8352

The scores were obtained through a 10-fold cross-validation using 80% of the data for training and 20% for testing. Both the training and testing datasets were stratified by morphological patterns. Not all morphological patterns were included as classification classes, however – as previously mentioned (see Table 4 in Section 3), the extraction from the lexicons (particularly srLex) resulted in many patterns that only occur once (approx. 62% of patterns in srLex and 49% of patterns in hrLex). These could not be part of a stratified sample, so they were excluded from the classification process.

It should also be noted that we evaluated the performance of the models based on their ability to correctly predict morphological pattern codes, not individual inflected forms. The scores could potentially be higher if taking

⁹ The models were trained using the scikit-learn library in Python (Pedregosa et al., 2011).

into account individual inflected forms – two morphological patterns with completely different pattern codes might in fact share a large number of inflected forms (e.g. patterns for animate or inanimate masculine common nouns, which only differ in the accusative singular form). It can also be difficult to predict the form of vocative singular of masculine nouns from the -a- declension (with the -a ending in the genitive singular form. For instance, unlike the hrLex example *abakuse* (see Table 2), similar nouns also exhibit vocative forms ending with -u: *dinosaurusu*, *glasu*, *fizikusu*, etc. (see Nikolić 2017). Another caveat is that hrLex and srLex are not gold-standard lexicons and were automatically generated, so the evaluations are not to be interpreted as comparisons to a manually annotated dataset, but rather how well the models represent the current state of the lexicons (described in more detail by Ljubešić et al., 2016).

Although the evaluation showed that the Linear Support Vector Classifier performed slightly better on srLex, we opted for Logistic Regression models in the end as that is also the same architecture used by the Slovene form generation models in *Pregibalnik*. In total, 24 final Logistic Regression models were trained in total (on the entire dataset). Their evaluations are shown in Table 7. It should also be noted that models were not developed for certain lexeme types that are not inflected and can be assigned a morphological pattern using a simple rule-based approach (e.g. interjections, conjunctions, abbreviations). The same rationale is applied to the Slovene form generator in *Pregibalnik*.

Table 7: Evaluation scores for Logistic Regression models for different lexeme types.

Lexeme type	hrLex 1.3			srLex 1.3		
	Accuracy	Baseline	F1-micro	Accuracy	Baseline	F1-micro
Ncm	0.65	0.28	0.85	0.64	0.27	0.85
Ncf	0.85	0.45	0.94	0.83	0.45	0.93
Ncn	0.97	0.81	0.99	0.93	0.77	0.95
Npm	0.86	0.43	0.99	0.86	0.43	0.99
Npf	0.93	0.88	0.93	0.92	0.88	0.93
Npn	0.57	0.33	0.85	0.56	0.33	0.85
Ag	0.58	0.33	0.85	0.57	0.33	0.84
Ap	0.99	0.99	0.99	0.93	0.92	0.93
As	0.99	0.99	0.99	0.99	0.99	0.99
Vm	0.70	0.21	0.96	0.64	0.19	0.97
Rg	0.72	0.73	0.72	0.71	0.73	0.72
Rr	0.95	0.86	0.96	0.87	0.81	0.96

All models achieve an above-baseline (majority classifier) accuracy with the exception of participial adjectives and general adverbs. A more detailed qualitative analysis is required to identify the exact root of this issue. However, it appears that in the current versions of both lexicons, many adverbs and adjectives seem to be lemmatized as infinitives of verbs (e.g. the participial adverb *abdicirajući* is tagged as an adverb, but lemmatized as the infinitive *abdicirati* 'to abdicate'; the same with *detonirajući* – *detonirati* 'to detonate' and *fermentirajući* – *fermentirati* 'to ferment'). This poses a problem because the lemma form is not present among the actual inflected forms, so the models probably do not learn much from lemma ending word parts. This lemmatization principle is also arguably counter-intuitive for users and introduces unnecessary ambiguities in the lexicon, which might cause more tagging errors if the tagger needs to decide between e.g. *abdicirati* as an adverb, adjective, or verb. This is something that can be addressed in future versions of the lexicons.

6. Preliminary Qualitative Evaluation

Due to space limitations, we only provide a brief preliminary manual evaluation of the performance of the models in this paper and leave a more detailed pattern-by-pattern analysis for future work.

Some classification errors can be attributed to inconsistencies in the lexicons. For instance, the proper masculine noun *Tomislavko* in hrLex features a morphological pattern with only singular forms, whereas the proper masculine noun *Žeško* features both singular and plural forms. The model correctly predicts the full morphological pattern in both cases. In some cases, the morphological pattern extraction revealed that several adjectives and adverbs in both lexicons feature incomplete patterns with only superlative forms, as is the case of *prevaran* 'deceitful'. The model correctly predicts the full pattern, so the classification can be partially used to identify inconsistencies and help with manual corrections.

On the other hand, there are several errors that can be expected due to inherent linguistic ambiguities. As in Slovene, Serbian and Croatian also have the distinction between animate and inanimate masculine nouns. Animacy is hard to predict for a simple model based simply on lemma-based features, so animate nouns are frequently misclassified as inanimate and vice versa. A similar issue occurs with adjectives and adverbs, for which the model has difficulties determining whether the pattern should feature gradation (with comparative and superlative forms) or not. These problems have also been encountered in Slovene models. In the future,

these issues will be addressed using post-processing methods that confirm morphological patterns with data in corpora, or large language models that can potentially fill the gaps of simple machine-learning models.

7. Conclusion and Future Work

In the paper, we presented the extraction of morphological patterns from the srLex and hrLex inflectional lexicons of Serbian and Croatian, and the development of open-access models for the automatic generation of inflected forms for Serbian and Croatian lexemes based on the MULTEXT-East morphosyntactic specifications. Both the extracted morphological patterns and the models are available on Github¹⁰ under the Creative Commons BY-SA 4.0 license.

The models can be used to expand the lexicon with additional lexemes from corpora. In our future work, we will implement the models into *Pregibalnik* to make them available as an API service. The extracted morphological patterns provide a good basis for a more thorough linguistic analysis, and the patterns can be further hierarchized (similar to Arhar Holdt & Čibej, 2018) and finally included as additional metadata into srLex and hrLex. Before manual validation, the patterns can be compared to the Apertium format patterns provided by Ljubešić et al. (2016) to identify similarities and discrepancies. Overall, the methodology to extract patterns and develop models is relatively language-independent and can also be applied to other languages (South Slavic or otherwise).

As a side-product, the analysis has also provided a list of potential inconsistencies in the existing version of the lexicons (e.g. the list of patterns occurring only once), which can be used in future manual validation campaigns to prioritize the most problematic lexemes.

Acknowledgment

The work presented in the paper was supported by the *COST Action CA21167 – Universality, Diversity, and Idiosyncrasy in Language Technology* (UniDive). The author also acknowledges the financial support from the Slovenian Research and Innovation Agency (research core funding No. P6-0411 – *Language Resources and Technologies for Slovene*). A sincere word of gratitude also goes to the anonymous reviewers for their constructive comments.

¹⁰ Github repository:

https://github.com/jakacibej/judig2024_morphological_inflection_srlex_hrlex

References

- [1] Arhar Holdt, Špela & Jaka Čibej. "Oblikoslovni vzorci v leksikonu Sloleks: izhodiščni nabor za samostalnike." *Slovenščina 2.0: Empirične, Aplikativne in Interdisciplinarne Raziskave*, 6(2). (2018): 33–66.
<https://doi.org/10.4312/slo2.0.2018.2.33-66>
- [2] Arhar Holdt, Špela. "Oblikoslovni vzorci za strojno procesiranje slovenščine." Arhar Holdt, Špela (ed.): *Nova slovnica sodobne standardne slovenščine: viri in metode*. Založba Univerze v Ljubljani. (2021): 87–124.
<https://doi.org/10.4312/9789610605478>
- [3] Čibej, Jaka, Kaja Gantar, Kaja Dobrovoljc, Simon Krek, Peter Holozan, Tomaž Erjavec, Miro Romih, Špela Arhar Holdt, Luka Krsnik, Marko Robnik-Šikonja. "Morphological lexicon Sloleks 3.0." Slovenian language resource repository CLARIN.SI, ISSN 2820-4042. (2022). <http://hdl.handle.net/11356/1745>.
- [4] Kosem, Iztok, Simon Krek, Polona Gantar. "Semantic data should no longer exist in isolation: the Digital Dictionary Database of Slovenian." *EURALEX XIX, Congress of the European Association for Lexicography, Lexicography for inclusion*. (2021): 81–83.
- [5] Krek, Simon, Špela Arhar Holdt, Tomaž Erjavec, Jaka Čibej, Andraž Repar, Polona Gantar, Nikola Ljubešić, Iztok Kosem, and Kaja Dobrovoljc. "Gigafida 2.0: The Reference Corpus of Written Standard Slovene." *Proceedings of the Twelfth Language Resources and Evaluation Conference, Marseille, France. European Language Resources Association*. (2020): 3340–3345.
- [6] Ljubešić, Nikola, Filip Klubička, Željko Agić, Ivo-Pavao Jazbec. "New Inflectional Lexicons and Training Corpora for Improved Morphosyntactic Annotation of Croatian and Serbian." *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Portorož, Slovenia. European Language Resources Association (ELRA). (2016): 4264–4270. <https://aclanthology.org/L16-1676/>
- [7] Ljubešić, Nikola, Filip Klubička. "Croatian web corpus hrWaC 2.1." Slovenian language resource repository CLARIN.SI, ISSN 2820-4042. (2016b). <http://hdl.handle.net/11356/1064>.
- [8] Ljubešić, Nikola, Filip Klubička. "Serbian web corpus srWaC 1.1." Slovenian language resource repository CLARIN.SI, ISSN 2820-4042. (2016a). <http://hdl.handle.net/11356/1063>.
- [9] Ljubešić, Nikola, Luka Terčon, Jaka Čibej. "The CLASSLA-Stanza model for morphosyntactic annotation of standard Slovenian 2.0." Slovenian language resource repository CLARIN.SI, ISSN 2820-4042. (2023). <http://hdl.handle.net/11356/1767>.

- [10] Ljubešić, Nikola. "*Inflectional lexicon hrLex 1.3*." Slovenian language resource repository CLARIN.SI, ISSN 2820-4042. (2019b).
<http://hdl.handle.net/11356/1232>
- [11] Ljubešić, Nikola. "*Inflectional lexicon srLex 1.3*." Slovenian language resource repository CLARIN.SI, ISSN 2820-4042. (2019a).
<http://hdl.handle.net/11356/1233>
- [12] Nikolić, Miroslav B. "Облици вокатива јединице именица мушког рода I врсте у српском књижевном језику". *Српски језик: студије српске и словенске*. - Vol. 22, No. 1 (2017): 5–34.
- [13] Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, vol. 12. (2011): 2825–2830.
<https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html>
- [14] Šnajder, Jan. "Models for predicting the inflectional paradigm of Croatian words." *Slovenščina 2.0, 1 (2)*, (2013): 1–34.
http://www.trojina.org/slovenscina2.0/arhiv/2013/2/Slo2.0_2013_2_02.pdf.
- [15] Terčon, Luka, Jaka Čibej, Nikola Ljubešić. "The CLASSLA-Stanza model for lemmatisation of standard Slovenian 2.0." Slovenian language resource repository CLARIN.SI, ISSN (2023): 2820–4042.
<http://hdl.handle.net/11356/1768>.

Modeli za automatsku morfološku fleksiju srpskog i hrvatskog jezika na osnovu morfoloških leksikona srLex i hrLex

Jaka Čibej

Sažetak

Mašinski čitljivi morfološki leksikoni otvorenog koda korisni su za morfosintaksičko označavanje korpusa i predstavljaju ključni korak ka sastavljanju savremenih baza podataka digitalnih rečnika. U radu predstavljamo prvi korak ka proširenju funkcionalnosti *Pregibalnika*, prilagođenog alata otvorenog koda za proširenje slovenačkog leksikona, tako da pokrije srpski i hrvatski jezik i pomoći će automatskom proširenju leksikona novim unosima. Opisujemo proces izdvajanja morfoloških obrazaca iz hrLex i srLex morfoloških leksikona hrvatskog i srpskog jezika, kao i robustan proces selekcije atributa na osnovu završnih delova reči. Atributi se koriste za razvoj serije modela mašinskog učenja za predviđanje morfoloških obrazaca za hrvatske i srpske lekseme, postižući prosečan F1-mikro rezultat od 0,85 (u zavisnosti od tipa lekseme). Ovo takođe pomaže da se identifikuju potencijalne nedoslednosti unutar trenutnih verzija leksikona. Izvučeni obrasci i modeli dostupni su pod licencom Creative Commons CC-BY 4.0.

Ključne reči: leksikon, morfologija, fleksija, proširenje leksikona, hrvatski, srpski

The ELEXIS-WSD Parallel Sense-Annotated Corpus and South Slavic Languages: Subcorpora for Croatian, Serbian, and Slovene

Scientific paper

DOI: 10.18485/judig.2025.1.ch3

*Jaka Čibej*¹,  0000-0002-3037-6848

*Ranka Stanković*²,  0000-0001-5123-6273

*Ana Ostroški Anić*³,  0000-0001-9999-0750

*Simon Krek*⁴,  0000-0001-8965-6863

*Carole Tiberius*⁵  0000-0002-7860-5427

Abstract

The open-source ELEXIS-WSD Parallel Sense-Annotated Corpus was developed within the ELEXIS project and in version 1.1 contains 2,024 sentences for each of 10 languages: Bulgarian, Danish, English, Spanish, Estonian, Hungarian, Italian, Dutch, Portuguese, and Slovene. Within the sentences, each content word (noun, adjective, verb, and adverb) has been assigned a corresponding sense from one of the 10 open-access sense inventories containing definitions. Within the context of the UniDive COST Action (CA21167), the corpus is being extended with several new languages, including South Slavic languages. In the paper, we focus on three subcorpora: Croatian, Serbian, and Slovene. We briefly describe the structure and purpose of the ELEXIS-WSD, then continue by describing

¹ Centre for Language Resources and Technologies, University of Ljubljana, Faculty of Arts, University of Ljubljana, jaka.cibej@ff.uni-lj.si

² University of Belgrade, Faculty of Mining and Geology, Chair for Mathematics and Informatics, ranka@rgf.rs

³ Department of General Linguistics, Institute for the Croatian Language, aostrosk@ihjj.hr

⁴ Artificial Intelligence Laboratory, “Jožef Stefan” Institute, Centre for Language Resources and Technologies, University of Ljubljana, simon.krek@ijs.si

⁵ Dutch Language Institute, Centre for Linguistics, carole.tiberius@ivdnt.org

the process of extending the corpus, which involves several different stages from translation to tokenization, lemmatization, and POS-tagging, to named entity and multiword expression/named entity annotation, and finally, word-sense disambiguation. We discuss some of the challenges encountered so far in these different phases with the corpus itself on the one hand, and the sense inventories on the other. We also describe the plans for future work on additional annotation layers within UniDive in order to further improve the ELEXIS-WSD corpus as a high-quality richly annotated manually curated dataset that is useful for NLP tasks such as word-sense disambiguation.

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

1. Introduction

The ELEXIS-WSD Parallel Sense-Annotated Corpus is a multilingual parallel corpus compiled within the ELEXIS project (European Lexicographic Infrastructure; 2017–2020)⁶ as part of a concerted effort to provide an open-source high-quality sense-annotated and linguistically informed dataset that can – among other things – be used for word-sense disambiguation tasks (hence the name ELEXIS-WSD). It is openly accessible as a dataset at the CLARIN.SI language repository (Martelli et al., 2023) – version 1.1 (the latest version at the time of writing this paper) is available under the Creative Commons BY-SA 4.0 license and covers 10 European languages: Bulgarian, Danish, Dutch, English, Estonian, Hungarian, Italian, Portuguese, Slovene, and Spanish.

The main incentive for the compilation of the ELEXIS-WSD corpus was the lack of high-quality manually curated sense-annotated datasets. During the ELEXIS project, existing datasets compiled for similar purposes were characterized by a number of drawbacks. First, the majority of them focused primarily on English (e.g., *SemCor* by Miller et al., 1993; datasets compiled within Senseval and SemEval shared tasks by Edmonds & Cotton, 2001; Snyder & Palmer, 2004; Navigli et al., 2007; Pradhan et al., 2007) or included only a limited number of instances in other languages (e.g., Agirre et al., 2010; Navigli et al., 2013; Moro & Navigli, 2015). Second, they relied on outdated or limited sense inventories or did not cover all part-of-speech categories. The endeavor within the ELEXIS project aimed to compile a resource that would link an annotated corpus with a high-quality sense inventory compiled through a lexicographic process (e.g. monolingual dictionaries).

⁶ The ELEXIS Project Website: <https://project.elex.is/>

The compilation process and structure of the corpus are described in more detail by Martelli et al. (2021). Each language subcorpus of ELEXIS-WSD contains the same 2,024 sentences with a total of approximately 35,000 tokens per language (a total of approximately 345,000 tokens). For each language, there is a separate sense inventory file containing senses, their IDs and definitions. Each content word in the corpus (adjectives, adverbs, verbs, and nouns according to the Universal Dependencies annotation scheme) is annotated with its corresponding sense from the sense inventory.

The purpose of the paper is to present the on-going work on extending the ELEXIS-WSD corpus with new languages and/or additional annotation layers, an effort that is being carried out within the CA21167 COST Action (UniDive). In this paper, we focus on the progress and challenges presented by the subcorpora of three South Slavic languages: Croatian, Serbian, and Slovene. We first describe the general goals and plans of the ELEXIS-WSD extension task within UniDive (Section 2), then proceed with describing the current state of the three subcorpora and the tasks that have been carried out so far (Section 3). We describe the plans for future developments within UniDive and conclude the paper with lessons learned and open questions for future work (Section 4).

2. Extension of ELEXIS-WSD within UniDive

The extension of ELEXIS-WSD is one of the tasks within the CA21167 COST Action titled *Universality, Diversity and Idiosyncrasy in Language Technology*⁷ (UniDive; 2022–2026), an international scientific network with the goal of reconciling language diversity with rapid progress in language technology. This provides the opportunity to extend the ELEXIS-WSD corpus in two ways: by adding other parallel subcorpora and thereby improving the corpus's coverage in terms of language diversity, and by adding additional annotation layers not present in the existing version, which will increase the usefulness of the corpus for a wide range of natural language processing tasks, as well as help develop and test out universal annotation guidelines and schemes across languages. The compilation process will be useful as a blueprint for potential similar datasets in the future, which will be crucial as evaluation benchmarks in the age of large language models.

Two elements need to be taken into account when adding a language to the ELEXIS-WSD corpus: the subcorpus (i.e. the set of 2,024 sentences)

⁷ UniDive Website: <https://unidive.lisn.upsaclay.fr/>

and the sense inventory (i.e. the set of senses and definitions with which the content words in the subcorpus are annotated).

At the time of writing this paper, a total of 8 new languages are included in the extension process: Greek, Romanian, Georgian, Macedonian, Polish, Ukrainian, Croatian, and Serbian. The ELEXIS-WSD corpus was initially compiled by extracting English sentences and their translation equivalents in other languages from the WikiMatrix dataset of parallel sentences (Schwenk et al., 2021) according to several criteria such as sentence length and number of polysemous words in the English sentence (see Martelli et al., 2021). Any missing translations were then translated manually. Within UniDive, we use English as a pivot language and start by first translating the sentences using a machine translation system (depending on the language), then manually correcting the result. The reason for skipping the sentence extraction from WikiMatrix for new languages within UniDive is that in the first version of the corpus, even with high-resource languages (such as Spanish), a significant amount of parallel sentences were missing, while some of those extracted were erroneous, poorly translated or machine-translated. Since we deal mostly with less-resourced languages in UniDive, we decided to use machine translations as a starting point.

The sense inventory which will be used to annotate tokens in the subcorpus should be in a machine-readable format (e.g. XML, TSV) and should ideally cover the majority of the approximately 5,000 ADV (adverb), ADJ (adjective), NOUN (common noun) and VERB (verb) lemmas from the ELEXIS-WSD corpus, along with their Universal Part-of-speech (UPOS) tags, the division of each lexeme into senses, and sense definitions. For instance, if the English lemma *premiere* appears as a verb in the subcorpus, the sense inventory should include all the available senses for *premiere* ~ VERB even if only one of the senses appears in the corpus. An example is shown in Table 1 to demonstrate the structure of the sense inventory of the English subcorpus, which is derived from the Open English WordNet (McCrae et al., 2019).

Table 1: Structure of the English sense inventory in ELEXIS-WSD 1.1.

Lemma	UPOS	ID ⁸	Definition
premiere	NOUN	...e1c-0	the first public performance of a play or movie
premiere	VERB	...e1d-0	be performed for the first time
premiere	VERB	...e1d-1	perform a work for the first time

⁸ The IDs in Table 1 were shortened for demonstration purposes.

The sense inventory should be available under the Creative Commons BY-SA 4.0 license and should preferably be based on lexicographic data (monolingual dictionaries or other lexicographic databases); because dictionaries are frequently not open-source and difficult to obtain, other semantic resources such as WordNets and Wiktionaries are acceptable as a fall-back.

After the machine translation phase, the process of extending the corpus involves several different annotation stages (such as tokenization, lemmatization, and POS-tagging). We describe these in more detail in the following sections using the examples of the subcorpora of South Slavic languages.

3. South Slavic Languages in ELEXIS-WSD

Currently, ELEXIS-WSD contains subcorpora for 5 South Slavic languages at various stages of development. Bulgarian and Slovene were included in the first phase of the compilation, so the initial annotation layers are complete. Bulgarian is currently not being developed further within UniDive, while Slovene continues with additional annotation layers (see Section 3.3). Macedonian, Croatian, and Serbian are new languages and started from the beginning within UniDive. Macedonian is in its early stages of editing machine translations, while Croatian (Section 3.1) and particularly Serbian (Section 3.2) have already advanced into subsequent stages.

3.1. Croatian Subcorpus – ELEXIS-WSD-hr

The Croatian subcorpus of ELEXIS-WSD started with the automatic translation and manual validation of English sentences. The sentences were translated into Croatian twice by two different systems: *Google Translate* and *Hrvojka*,⁹ a custom-made Croatian national machine-translation platform developed within the NLPT project for the benefits of public administration, the economic and academic communities. The Croatian UniDive team took parallel machine-translations and compared them to the English original to provide the final manually-curated Croatian version of each sentence.

The challenges with translation stemmed mostly from the nature of the sentences included in the corpus. The sentences are unrelated and disjointed as they are extracted from different Wikipedia texts. Their content is frequently very encyclopaedic and includes specific terms from

⁹ *Hrvojka* is available as part of the Croatian National Language Technology Platform: <https://hrvojka.gov.hr/>

different scientific areas. In some cases, the sentences are very short and contain some unclear references (such as pronouns), for which gender and number sometimes had to be disambiguated. This was done by either consulting other subcorpora within ELEXIS-WSD or searching for the context of the original sentences on Wikipedia.

The translations have recently been finished and the subcorpus will proceed with the next annotation phases, which include tokenization, lemmatization, and part-of-speech tagging. This will be done using Universal Dependencies 2.15 tagging models available in *UDPipe 2*,¹⁰ specifically the *croatian-set-ud-2.15-241121* model. The layers will then be manually corrected in the open-source INCEpTION annotation platform (Klie et al., 2018) before proceeding to more advanced annotations such as word-sense disambiguation (see Section 4).

As for the sense inventory that will be used to annotate senses in the subcorpus, ELEXIS-WSD-hr will use the *School Dictionary of Croatian* (*Školski rječnik hrvatskoga jezika*; Birtić et al., 2012),¹¹ a normative dictionary that contains approximately 31,000 entries and is intended for Croatian students in the final years of primary and secondary schools. Beside grammatical and morphological information on entries (part-of-speech, aspect, gender, etc.) and their sense definitions, the dictionary also includes multi-word expressions, synonyms, antonyms, and normative labels. However, additional work is required to prepare the sense inventory. While the dictionary is available in an XML format, the data is unstructured, with different types of data (from morphological information to definitions and multiword expressions) contained in a single string that is formatted according to lexicographic guidelines. Using rule-based approaches and regular expressions, the dictionary is currently being converted into the appropriate format (see Table 1). This is a work-in-progress that will require some additional manual verification. The initial phases of the process have been described by Runjaić & Čibej (2025). Any lemmas from ELEXIS-WSD-hr that are absent in the dictionary will be manually added to the sense inventory along with their sense divisions and definitions.

3.2. Serbian Subcorpus – ELEXIS-WSD-sr

Unlike Croatian, the sentences for the Serbian subcorpus (in Latin script) were translated only with Google Translate, then manually validated by 8 Serbian native speakers (Krstev et al., 2024a) in order to avoid literal

¹⁰ UDPipe 2: <https://lindat.mff.cuni.cz/services/udpipe/>

¹¹ *The School Dictionary of Croatian* is also available in an online interface: <https://rjecnik.hr/>

or incorrect translations (particularly in the case of multi-word expressions) and to resolve issues already mentioned in Section 3.1 (resolution of ambiguous gender features in pronouns, specific terminology, etc.), a particular challenge for Serbian in this phase was the presence of a large number of foreign proper nouns and named entities, which according to Serbian orthographic rules should not be written in their original form, but transliterated phonetically. This was again validated by two people.

At the time of writing this paper, several annotation layers (tokenization, lemmatization, POS-tagging, as well as named entity recognition) have already been completed, while others are still in progress. Tokenization, lemmatization, and POS-tagging were first annotated automatically (Stanković et al., 2022) then checked by at least three evaluators. Named entities (NEs) were automatically annotated with the *sr_ner_tesla_j355* model (Ikonić Nešić et al., 2024) using a schema of 7 categories; an eighth category – PRODUCT – was introduced manually. In total, 2,294 NEs were manually verified: PERS (439), LOC (710), ORG (330), DEMO (301), ROLE (309), EVENT (60), WORK (75), PRODUCT (80). Named entities were manually linked with Wikidata using the INCEpTION platform (Klie et al., 2018). As a result, 1,949 NEs were linked with the existing Wikidata items, which leaves 345 unlinked NEs. The corpus was also automatically annotated with verbal and nominal MWEs (as elaborated in Krstev et al., 2024b), which was followed by manual validation. The classification of verbal MWEs followed the PARSEME guidelines,¹² while non-verbal MWEs will be categorized according to the extended PARSEME guidelines that are being developed within UniDive (more on this in Section 4).

Because there is no open-access digital descriptive dictionary of Serbian, the Serbian sense inventory for ELEXIS-WSD-sr is based on the *Serbian WordNet* (SrpWN; Stanković et al., 2018). However, SrpWN does not cover all the lemmas from the corpus and their senses, and needed to be expanded with the missing synsets. The inventory for the English subcorpus is derived from Open English WordNet, so its 13,703 synsets were aligned with SrpWN, resulting in 5,997 matches. A comparison with ELEXIS-WSD-en sense annotations highlighted 2,130 missing synsets that required urgent addition. The synonyms and definitions from Open English Wordnet were automatically translated using Google API and OpenAI and were later post-edited. For the rest of content words without definitions in the Serbian sense inventory, the entries were added by prompt engineering

¹² PARSEME Guidelines for MWE Annotation: <https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.3/>

using Gemini (gemini-2.0-flash-exp) and ChatGPT (gpt-4), followed by post-editing. The Serbian inventory is currently being expanded with definitions of identified MWEs, mostly manually by consulting various resources.

The demo version of the current state of the annotated corpus is available as NIF (NLP Interchange Format) corpus.¹³ Once completed, the ELEXIS-WSD-sr corpus will be the first sense-annotated corpus using the Serbian WordNet.

3.3. Slovene Subcorpus – ELEXIS-WSD-sl

As one of subcorpora from the original version of ELEXIS-WSD, ELEXIS-WSD-sl already has manually checked tokenization, lemmatization, part-of-speech tags, and sense annotations. Because no open-source Slovene dictionary was available at the time and the existing version of the Slovene WordNet – *slowNet 3.1* (Fišer, 2015) – lacked a significant number of Slovene definitions and thus provided poor coverage for sense annotation, the sense inventory was compiled from scratch by collocating the activity with ongoing efforts to develop the *Digital Dictionary Database of Slovene* (DDDS; Kosem et al., 2021). First, a frequency list of lemmas and their part-of-speech tags was extracted from ELEXIS-WSD-sl. For each lemma, a sense division was determined by lexicographers using both ELEXIS-WSD-sl and other Slovene corpora (such as the *Gigafida Corpus of Written Standard Slovene*; Krek et al., 2019). Individual occurrences of lemmas were then assigned appropriate senses from the DDS, taking into account multi-word expressions (MWEs) as well – each individual token pertaining to a MWE was annotated with the sense of the entire expression.

While some information on MWEs can be extrapolated from sense annotations, the subcorpus does not yet explicitly contain multiword expression annotations. Within UniDive, however, all MWEs in ELEXIS-WSD-sl will be annotated with categories proposed first in the PARSEME COST Action for verbal MWEs (such as light-verb constructions and verbal idioms; see e.g. Savary et al., 2023) and further developed within UniDive for non-verbal MWEs (nominal, adjectival, adverbial, and functional MWEs; an initial taxonomy was proposed by Ramisch (2023)).

Another annotation layer concerns dependency relations according to the Universal Dependencies annotation scheme (for a description of the plans to add UD syntax annotations to ELEXIS-WSD, see Tiberius et al.

¹³ The ELEXIS-WSD-sr corpus as NIF is available at endpoint: <http://fuseki.jerteh.rs/#/dataset/sr-ELEXIS-UNIDIVE-demo/query>

(2024); the plan is to have every language annotated with at least automatically assigned (or, ideally, manually validated) dependency relations). The sentences from ELEXIS-WSD-sl were also included in the *SUK Training Corpus of Written Slovene* (Arhar Holdt et al., 2024), where they were manually annotated with dependency relations. However, in the process, some changes occurred in morphosyntactic tagging as well, so the mapping process between the SUK and ELEXIS-WSD-sl versions is not entirely straightforward. The discrepancies are currently being resolved.

4. Future Plans

In addition to the annotation layers already mentioned (lemmatization, morphosyntactic tags, sense annotations, UD dependency relations, and MWE annotations according to the PARSEME/UniDive categorization), the finalized versions of the subcorpora in ELEXIS-WSD will also contain (at least) two additional annotation layers.

The first layer concerns named entities (NEs). For the English subcorpus, NEs have already been annotated, but the annotations have not yet been implemented in version 1.1 as they were only meant as a reference for sense annotation. In addition, only spans were annotated. In the Serbian subcorpus, however, NEs have been annotated much more systematically (Krstev et al., 2024a) using a set of 8 classes. There is still the open question of the set of categories used for annotation and whether the set is universal enough to be applied to each language within ELEXIS-WSD (see Krstev et al., 2024b). Unlike the PARSEME guidelines for categorizing MWEs, NEs currently have no universal guidelines within UniDive yet (an example of annotating a corpus with both MWEs and NEs is provided by e.g. Candito et al. 2020). Once the annotation scheme is complete, the NE annotations will be propagated across all subcorpora. In addition, the annotated NEs will be further enriched by links to their corresponding WikiData instances.

The second layer is an additional semantic layer alongside the existing sense annotations that link individual tokens to senses in the sense inventory. Each sense annotation will also be annotated with a *supersense*, a more coarse-grained semantic category denoting broad concepts such as ANIMAL, ARTIFACT, FEELING, GROUP, and PLANT. The set of supersenses are used to more broadly categorize groups of semantically similar synsets in WordNets (Jurafsky & Martin, 2025). There are a total of 44 supersense categories (26 for nouns, 15 for verbs, 2 for adjectives, and 1 for adverbs). Because a number of subcorpora in ELEXIS-WSD already use different WordNets as sense inventories, the propagation of supersense

annotations from synsets onto individual sense-annotated tokens is a trivial task. Propagation to other subcorpora which use non-WordNet sense inventories will require some more work, but to an extent, the process can be automated with large language models to reduce the amount of manual validation required.

5. Conclusion

We have briefly presented the on-going effort of extending the ELEXIS-WSD Parallel Sense-Annotated Corpus with two new South Slavic languages (Croatian and Serbian), as well as additional work on the existing Slovene subcorpus. We described some of the challenges encountered in different annotation phases of the corpus, as well as different solutions to providing an open-access sense inventory for corpus annotation.

Just like version 1.1, all the data produced will be openly accessible at the CLARIN.SI language repository under the Creative Commons BY-SA 4.0 licence. The sense inventories will also be uploaded to WikiBase¹⁴ and will be available for querying at a SPARQL endpoint.

The work within UniDive will result in a richly annotated multi-layered corpus that can be useful for NLP tasks or contrastive linguistic analyses. The experience of compiling the dataset will be invaluable in the future as similar resources are developed as benchmark datasets to evaluate the performance of large language models for complex semantic tasks.

Acknowledgment

The work presented in the paper was supported by the *COST Action CA21167 – Universality, Diversity and Idiosyncrasy in Language Technology* (UniDive). The authors also acknowledge the financial support from the Slovenian Research and Innovation Agency (research core funding No. P6-0411 – *Language Resources and Technologies for Slovene*). A sincere word of gratitude also goes to the anonymous reviewers for their insightful comments.

¹⁴ Wikibase: <https://wikiba.se/>

References

- [1] Agirre, Eneko, Oier Lopez de Lacalle, Christiane Fellbaum, Shu-Kai Hsieh, Maurizio Tesconi, Monica Monachini, Piek Vossen and Roxanne Segers. "SemEval-2010 Task 17: All-Words Word Sense Disambiguation on a Specific Domain". *Proceedings of the 5th International Workshop on Semantic Evaluation*. Uppsala, Sweden. Association for Computational Linguistics. (2010): 75–80. <https://aclanthology.org/S10-1013/>
- [2] Arhar Holdt, Špela, Simon Krek, Kaja Dobrovoljc, Tomaž Erjavec, Polona Gantar, Jaka Čibej, Eva Pori, Luka Terčon, Tina Munda, Slavko Žitnik, Nejc Robida, Neli Blagus, Sara Može, Nina Ledinek, Nanika Holz, Katja Zupan, Taja Kuzman, Teja Kavčič, Iza Škrjanec, Dafne Marko, Lucija Jezeršek, Anja Zajc, 2024, Training corpus SUK 1.1, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, <http://hdl.handle.net/11356/1959>.
- [3] Birtić, Matea, Goranka Blagus Bartolec, Lana Hudeček, Ljiljana Jojić, Barbara Kovačević, Kristian Lewis, Ivana Matas Ivanković, Milica Mihaljević, Irena Miloš, Ermina Ramadanović and Domagoj Vidović. "*Školski rječnik hrvatskoga jezika*". Školska knjiga, Zagreb, Croatia. (2012)
- [4] Marie Candito, Mathieu Constant, Carlos Ramisch, Agata Savary, Bruno Guillaume, Yannick Parmentier, Silvio Ricardo Cordeiro. "A French corpus annotated for multiword expressions and named entities". *Journal of Language Modelling*, 8(2). (2020): 415–479. <https://doi.org/10.15398/jlm.v8i2.265>
- [5] Edmonds, Philip and Scott Cotton. "SENSEVAL-2: Overview". *Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems*. Toulouse, France. Association for Computational Linguistics. (2001): 1–5. <https://aclanthology.org/S01-1001/>
- [6] Fišer, Darja. "Semantic lexicon of Slovene sloWNet 3.1". Slovenian language resource repository CLARIN.SI, ISSN 2820-4042. (2015) <http://hdl.handle.net/11356/1026>
- [7] Ikonić Nešić, Milica, Saša Petalinkar, Mihailo Škorić, Ranka Stanković. "BERT Downstream Task Analysis: Named Entity Recognition in Serbian". In: Trajanović, M., Filipović, N., Zdravković, M. (eds) *Disruptive Information Technologies for a Smart Society. ICIST 2024. Lecture Notes in Networks and Systems*, vol 860. Springer, pp 333–347 Cham. (2024) https://doi.org/10.1007/978-3-031-71419-1_29
- [8] Jurafsky, Daniel and James H. Martin. "Word Senses and WordNet. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models". (2025). <https://web.stanford.edu/~jurafsky/slp3/G.pdf>
- [9] Klie, Jan-Christoph, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho and Iryna Gurevych. "The INCEpTION Platform: Machine-Assisted and

- Knowledge-Oriented Interactive Annotation". *Proceedings of System Demonstrations of the 27th International Conference on Computational Linguistics (COLING 2018)*, Santa Fe, New Mexico, USA (2018): 5–9.
<https://aclanthology.org/C18-2002/>
- [10] Kosem, Iztok, Simon Krek and Polona Gantar. "Semantic data should no longer exist in isolation: the digital dictionary database of Slovenian." *Proceedings of the XIX EURALEX International Congress: Lexicography for Inclusion*. Komotini: SynMorPhoSe Lab, Democritus University of Thrace. (2021): 81–83. https://elex.is/wp-content/uploads/2021/09/Semantic-Data-should-no-longer-exist-in-isolation-the-Digital-Dictionary-Database-of-Slovenian_Kosem-Krek-Gantar_EURALEX2020.pdf
- [11] Krek, Simon, Carole Tiberius, Kaja Dobrovoljc, Jaka Čibej, Polona Gantar, Jelena Kallas, Kristina Koppel, Svetla Peneva Koeva, Veronika Lipp, László Simon: "The ELEXIS Parallel Sense-annotated Corpus". *UniDive 1st General Meeting*. Paris, France. (2023).
https://unidive.lisn.upsaclay.fr/lib/exe/fetch.php?media=meetings:2023-saclay:abstracts:45_krek_et_al_the_elexis_parallel_sense_annotated_corpus.pdf
- [12] Krek, Simon, Tomaž Erjavec, Andraž Repar, Jaka Čibej, Špela Arhar Holdt, Polona Gantar, Iztok Kosem, Marko Robnik-Šikonja, Nikola Ljubešić, Kaja Dobrovoljc, Cyprian Laskowski, Miha Grčar, Peter Holozan, Simon Šuster, Vojko Gorjanc, Marko Stabej, Nataša Logar. "Corpus of Written Standard Slovene Gigafida 2.0". *Slovenian language resource repository CLARIN.SI*, ISSN 2820-4042. (2019) <http://hdl.handle.net/11356/1320>.
- [13] Krstev, Cvetana, Aleksandra Marković, Ranka Stanković. "Annotation of MWEs and NEs in the Serbian extension of ELEXIS-WSD: comparisons, solutions and open questions." *Unidive 2nd General Meeting*. Naples, Italy. (2024b)
https://unidive.lisn.upsaclay.fr/lib/exe/fetch.php?media=meetings:general_meetings:2nd_unidive_general_meeting:17_annotation_of_mwes_and_nes_in_.pdf
- [14] Krstev, Cvetana, Aleksandra Marković, Ranka Stanković, Milica Ikonić Nešić. "Progress in SR-ELEXIS Semantic Annotation: Focusing on Multiword Expressions, Named Entities, and Sense Repository." *Unidive 3rd General Meeting*. Budapest, Hungary. (2025)
https://unidive.lisn.upsaclay.fr/lib/exe/fetch.php?media=meetings:general_meetings:3rd_unidive_general_meeting:32_progress_in_sr_elexis_seman.pdf
- [15] Krstev, Cvetana, Bojana Djordjević, Sanja Antonić, Nevena Ivković-Berček, Zorica Zorica, Vesna Crnogorac, and Ljiljana Macura. "Cooperative Work in Further Development of Serbian WordNet". *INFOteka, IX(12)*. (2008): 59–78.
- [16] Krstev, Cvetana, Ranka Stanković, Aleksandra Marković, Teodora Mihajlov. "Towards the semantic annotation of SR-ELEXIS corpus: Insights into Multiword Expressions and Named Entities". *Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD 2024)*, SIGLEX, ACL,

- UniDive CA21167. eds.: Bhatia, Archana et al. virtual presentation, May 25, 2024. (2024a): 74-84. <https://aclanthology.org/2024.mwe-1.15.pdf>
- [17] Martelli, Federico, Roberto Navigli, Simon Krek, Jelena Kallas, Polona Gantar, Svetla Koeva, Sanni Nimb, Bolette Sandford Pedersen, Sussi Olsen, Margit Langemets, Kristina Koppel, Tiiu Üksik, Kaja Dobrovoljc, Rafel Ureña-Ruiz, José-Luis Sancho-Sánchez, Veronika Lipp, Tamás Váradi, András Györffy, László Simon, Valeria Quochi, Monica Monachini, Francesca Frontini, Carole Tiberius, Rob Tempelaars, Rute Costa, Ana Salgado, Jaka Čibej, Tina Munda, Iztok Kosem, Rebeka Roblek, Urška Kamenšek, Petra Zaranšek, Karolina Zgaga, Primož Ponikvar, Luka Terčon, Jonas Jensen, Ida Flörke, Henrik Lorentzen, Thomas Troelsgård, Diana Blagoeva, Dimitar Hristov, Sia Kolkovska: "Parallel sense-annotated corpus ELEXIS-WSD 1.1". *Slovenian language resource repository CLARIN.SI*, ISSN 2820-4042. (2023). <http://hdl.handle.net/11356/1842>
- [18] Martelli, Federico, Roberto Navigli, Simon Krek, Jelena Kallas, Polona Gantar, Svetla Koeva, Sanni Nimb, Bolette Pedersen, Sussi Olsen, Margit Langemets, Kristina Koppel, Tiiu Üksik, Kaja Dobrovoljc, Rafael Ureña Ruiz, José Luis Sancho Sánchez, Veronika Lipp, Tamás Váradi, András Györffy, Simon László and Tina Munda. "Designing the ELEXIS Parallel Sense-annotated Dataset in 10 European Languages." *Electronic lexicography in the 21st century. Proceedings of the eLex 2021 conference*. (2021): 377–395. https://elex.link/elex2021/wp-content/uploads/2021/08/eLex_2021_22_pp377-395.pdf
- [19] McCrae, John P., Alexandre Rademaker, Francis Bond, Ewa Rudnicka and Christiane Fellbaum. "English WordNet 2019 – An Open-Source WordNet for English". *Proceedings of the 10th Global WordNet Conference – GWC 2019, Wrocław*. (2019): 245–252. <https://aclanthology.org/2019.gwc-1.31/>
- [20] Miller, George A., Claudia Leacock, Randee Teng, Ross T. Bunker. "A semantic concordance." *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*. (1993) <https://aclanthology.org/H93-1061/>
- [21] Moro, Andrea and Roberto Navigli. "SemEval-2015 Task 13: Multilingual All-Words Sense Disambiguation and Entity Linking". *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. Denver, Colorado. Association for Computational Linguistics. (2015): 288–297. <https://aclanthology.org/S15-2049/>
- [22] Navigli, Roberto, Kenneth C. Litkowski, and Orin Hargraves. "SemEval-2007 Task 07: Coarse-Grained English All-Words Task". *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*. Prague, Czech Republic. Association for Computational Linguistics. (2007): 30–35. <https://aclanthology.org/S07-1006/>

- [23] Navigli, Roberto, David Jurgens, and Daniele Vannella. "SemEval-2013 Task 12: Multilingual Word Sense Disambiguation". *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. Atlanta, Georgia, USA. Association for Computational Linguistics. (2013): 222–231. <https://aclanthology.org/S13-2040/>
- [24] Pradhan, Sameer, Edward Loper, Dmitriy Dligach, and Martha Palmer. "SemEval-2007 Task-17: English Lexical Sample, SRL and All Words". *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*. Prague, Czech Republic. Association for Computational Linguistics. (2007): 87–92. <https://aclanthology.org/S07-1016/>
- [25] Ramisch, Carlos. "A taxonomy proposal for multiword expressions". *2nd UniDive General Meeting*. Naples, Italy. (2024) https://unidive.lisn.upsaclay.fr/lib/exe/fetch.php?media=meetings:general_meetings:2nd_unidive_general_meeting:23_a_taxonomy_proposal_for_multiw.pdf
- [26] Runjaić, Siniša and Jaka Čibej. "Conversion of the School Dictionary of Croatian into a Sense Inventory for the ELEXIS-WSD Parallel Sense-Annotated Corpus: A Case Study". *UniDive 3rd General Meeting*. Budapest, Hungary. (2025) https://unidive.lisn.upsaclay.fr/lib/exe/fetch.php?media=meetings:general_meetings:3rd_unidive_general_meeting:20_conversion_of_the_school_di.pdf
- [27] Savary, Agata, Sara Stymne, Verginica Barbu Mititelu, Nathan Schneider, Carlos Ramisch, Joakim Nivre. "PARSEME Meets Universal Dependencies: Getting on the Same Page in Representing Multiword Expressions". *Northern European Journal of Language Technology*, Vol. 9, No. 1. (2023). <https://nejlt.ep.liu.se/article/view/4453>
- [28] Schwenk, Holger, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. "WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia". *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. (2021): 1351–1361. <https://aclanthology.org/2021.eacl-main.115/>
- [29] Snyder, Benjamin and Martha Palmer. "The English all-words task". *Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*. Barcelona, Spain. Association for Computational Linguistics. (2004): 41–43. <https://aclanthology.org/W04-0811/>
- [30] Stanković, Ranka, Miljana Mladenović, Ivan Obradović, Marko Vitas, and Cvetana Krstev. "Resource-based WordNet Augmentation and Enrichment". *Proceedings of the Third International Conference Computational Linguistics in Bulgaria (CLIB 2018)*, pages 104–114, Sofia, Bulgaria. Institute for Bulgarian Language "Prof. Lyubomir Andreychin", Bulgarian Academy of Sciences. (2018)

- [31] Stanković, Ranka, Mihailo Škorić, and Branislava Šandrih Todorović. 2022. "Parallel Bidirectionally Pretrained Taggers as Feature Generators" *Applied Sciences* 12, no. 10: 5028. <https://doi.org/10.3390/app12105028>
- [32] Tiberius, Carole, Jaka Čibej, Jelena Kallas, Kertu Saul, Kadri Muischnek, Simon Krek. "UD Syntax for the ELEXIS-WSD Parallel Sense-Annotated Corpus: A Pilot Study". *UniDive 2nd General Meeting*. Naples, Italy. (2024) https://unidive.lisn.upsaclay.fr/lib/exe/fetch.php?media=meetings:general_meetings:2nd_unidive_general_meeting:31_ud_syntax_for_the_elexis_paral.pdf

ELEXIS-WSD paralelni semantički anotirani korpus i južnoslovenski jezici: potkorpusi za hrvatski, srpski i slovenački

Jaka Čibej, Ranka Stanković, Ana Ostroški Anić, Simon Krek, Carole Tiberius

Sažetak

Otvoreni paralelni semantički anotiran korpus ELEXIS-WSD razvijen je u okviru projekta ELEXIS i u verziji 1.1 sadrži 2.024 rečenice za svaki od 10 jezika: bugarski, danski, engleski, španski, estonski, mađarski, italijanski, holandski, portugalski i slovenački. U rečenicama, svakoj reči koja nose značenje (imenice, pridevi, glagoli i prilozi) dodeljeno je odgovarajuće značenje iz jednog od 10 otvorenih repozitorijuma značenja koji sadrže definicije. U kontekstu UniDive COST akcije (CA21167), korpus se proširuje sa nekoliko novih jezika, uključujući južnoslovenske jezike. U ovom radu fokusiramo se na tri potkorpuse: hrvatski, srpski i slovenački. Ukratko opisujemo strukturu i namenu ELEXIS-WSD korpusa, a zatim nastavljamo sa opisom procesa proširenja korpusa, koji uključuje nekoliko različitih faza – od prevođenja do tokenizacije, lematizacije i obeležavanja vrsta reči (POS-tagging), zatim anotacije imenovanih entiteta i polileksemskih izraza/imenovanih entiteta, i konačno, razrešavanja višeznačnosti reči. U radu razmatramo neke od izazova sa kojima smo se do sada susreli u različitim fazama, kako u pripremi korpusa, tako i u pripremi repozitorijuma značenja. Takođe opisujemo planove za budući rad na dodatnim slojevima anotacije u okviru UniDive projekta, sa ciljem daljeg poboljšanja ELEXIS-WSD korpusa kao visokokvalitetnog, bogato anotiranog i ručno verifikovanog skupa podataka koji je koristan za NLP zadatke kao što je razrešavanje višeznačnosti reči.

Ključne reči: semantička anotacija, paralelni korpus, značenja, južnoslovenski jezici, slovenački, hrvatski, srpski

Značaj digitalnog korpusa i jezičkih alata u jezičkoj analizi forenzičkih tekstova na srpskom jeziku

Naučni rad

DOI: 10.18485/judig.2025.1.ch4

Jelena Redli¹  0000-0003-2518-5286

Apstrakt

Forenzička lingvistika (FL) je interdisciplinarna oblast koja kombinuje jezičke, pravne i digitalne tehnologije radi boljeg razumevanja jezičkih fenomena u pravnom i kriminalističkom kontekstu. U vidu pokušaja utvrđivanja autorstva FL se primenjuje od 18. veka još od prvih rasprava o autorstvu Biblije da bi se kasnije proširila na otkrivanje potencijalnih kriminalnih aktivnosti i sl. Za sve FL aktivnosti ključni su specijalizovani korpusi koji služe za obuku sistema tokom razvoja. Iako je upotreba ovih korpusa nezaobilazna u globalnom kontekstu, srpski jezik ostaje nedovoljno istražen zbog nedostatka specijalizovanih korpusa forenzičkih tekstova.

Ovaj rad istražuje potencijal kreiranja i primene korpusa forenzičkih tekstova na srpskom jeziku, fokusirajući se na njihovu ulogu u analizi autentičnosti teksta, identifikaciji autora i rešavanju pravnih sporova putem jezičkih dokaza. Metodologija koju rad predlaže obuhvata različite vrste forenzičkih tekstova kao što su policijski izveštaji, pravni dokumenti, preteće poruke i oprostajna pisma. Takođe se razmatraju napredni jezički alati za automatsku analizu teksta, kao i dubinsku analizu leksičke frekvencije, sintakse, stilskih markera i jezičkih obrazaca.

Pored teorijskog doprinosa, rad demonstrira praktičnu primenu korpusa u srpskom kriminalističko-pravnom kontekstu, ističući potencijalne izazove u razvoju kao što su tehnički, pravni i etički aspekti.

¹ Filozofski fakultet, Univerzitet u Novom Sadu, redli@ff.uns.ac.rs

Poseban naglasak stavlja se na poštovanje etičkih i pravnih smernica tokom prikupljanja i obrade osetljivih podataka.

Na kraju, rad pruža preporuke za buduća istraživanja i implementaciju, naglašavajući važnost interdisciplinarnе saradnje između lingvista, pravnika, policije i IT stručnjaka u razvoju jezičkih resursa za forenzička istraživanja u Srbiji.

Ključne reči: forenzička lingvistika, srpski jezik, pravni jezik, digitalni korpus, jezički alati, forenzička analiza teksta, identifikacija autora.

1. Uvod

Forenzička lingvistika (FL) relativno je mlada disciplina koja se bavi analizom jezika u pravnim i kriminalističkim kontekstima. Njeni zvanični počeci datiraju iz 1968. godine, kada je švedski lingvista Jan Svartvik po prvi put primenio lingvističke metode u analizi izjava datih policiji u slučaju *Evans* (Svartvik, 1968). Forenzička lingvistika obuhvata širok spektar metoda i tehnika za analizu tekstova i govora u pravnim okvirima, s posebnim fokusom na jezičke dokaze koji se koriste u pravnim postupcima. Ona kombinuje lingvistička znanja s pravnim i kriminalističkim metodologijama kako bi se istražili složeni jezički fenomeni prisutni u forenzičkim tekstovima, kao što su izjave osumnjičenih, svedoka, policijski izveštaji, preteće poruke i drugi relevantni dokumenti.

Jedna od glavnih uloga FL jeste identifikacija autora spornih tekstova, analiza autentičnosti izjava, te evaluacija pravnih dokaza putem jezičke analize. Upotrebom metoda iz oblasti fonetike, leksikologije, sintakse, semantike, pragmatike, analize diskursa i lingvistike teksta forenzički lingvisti istražuju jezičke karakteristike tekstova u cilju utvrđivanja autorstva, verodostojnosti i drugih karakteristika datih izjava koje mogu poslužiti kao deo dokaza, a ponekad i kao jedini dokaz (Coulthard, 2005, p. 10; Olsson, 2010, p. 11).

Iako su istraživanja upotrebe digitalnih korpusa i alata u analizi forenzičkih tekstova na srpskom jeziku dugo bila ograničena, poslednjih godina zabeležen je značajan napredak (v. Vorkapić et al., 2017), pri čemu se u obradi jezika primenjuju kako tradicionalni, tako i savremeni alati, uključujući velike jezičke modele. Nedostatak specijalizovanih digitalnih korpusa forenzičkih tekstova na srpskom jeziku značajno ograničava mogućnosti za detaljnu analizu i praktičnu primenu jezičkih alata. U razvijenim pravnim sistemima, digitalni korpusi postaju važni alati za identifikaciju autora, analizu sintakse, leksike i stilskih markera u jeziku osumnjičenog ili nepoznatog autora nekog forenzičkog teksta.

Na srpskom jeziku, dosad nije kreiran specijalizovan korpus forenzičkih tekstova, što predstavlja velik izazov u kriminalističko-pravnom kontekstu. Stoga postoji potreba za razvojem korpusa koji bi uključivao različite vrste tekstova relevantnih za FL.

Zato je glavni cilj ovog rada da istraži značaj digitalnog korpusa i jezičkih alata u forenzičkoj analizi tekstova na srpskom jeziku. Kreiranje specijalizovanog korpusa omogućilo bi lakšu identifikaciju autora, analizu autentičnosti tekstova i unapređenje postupaka u pravnim sporovima. Ovaj rad pruža metodološki okvir za formiranje takvog korpusa i daje smernice za njegovu praktičnu primenu u pravnim i kriminalističkim istragama. Time bi srpski jezik postao deo globalne mreže forenzičkih istraživanja, uz poštovanje etičkih i pravnih smernica za obradu osetljivih podataka.

2. Pregled literature

2.1. Korpusna lingvistika i forenzička lingvistika

Korpusna lingvistika je postala temeljna metoda za analizu jezika u različitim oblastima, uključujući i FL. Razvoj velikih, digitalnih zbirki tekstova omogućio je lingvistima da proučavaju jezičke fenomene na način koji ranije nije bio moguć. S. Blackwell, jedno od prominentnih imena u ovom polju, ističe da, iako je korpusna lingvistika prisutna od šezdesetih godina XX veka, njena primena u FL relativno je nova (Blackwell, 2009, p. 5). Korpusni alati omogućavaju preciznu analizu sintakse, leksike i frazeologije, što pomaže u identifikaciji autora i analizi autentičnosti spornih tekstova poput pretećih pisama ili anonimnih poruka.

Jedan od glavnih problema u ranim radovima na polju FL bio je korišćenje opštih korpusa kao što je Sinklerov *Bank of English*² umesto specifičnih forenzičkih korpusa, što je često doводilo do nejasnih rezultata (Ibid.). Korpusna lingvistika i forenzička lingvistika sada se sve više povezuju, kako Kultard naglašava, posebno u analizi pravnih i kriminalističkih tekstova (Coulthard, 2013, p. 202). Njegova istraživanja pokazala su značaj specijalizovanih korpusa za analizu stilskih markera u pravnim tekstovima, čime se povećava tačnost identifikacije autora i verodostojnosti dokumenata.

² Bank of English je danas dostupan preko CQPWeb platforme, na serveru Univerziteta u Birminghamu, što omogućava nastavnom osoblju i studentima ovog univerziteta da pristupe ovom i drugim korpusima kroz jedinstven interfejs. Pristup sistemu zahteva registraciju putem imejl adrese sa domenom bham.ac.uk. (v. University of Birmingham CQPWeb server: <http://cqpweb.bham.ac.uk>).

2.2. Primena korpusa u forenzičkim analizama

Korpusi su ključni u forenzičkim analizama zbog svoje sposobnosti da pruže uvid u frekvenciju pojavljivanja određenih jezičkih elemenata u tekstu. Na primer, istraživanja Dejvida Vrajta o n-gramima pokazala su kako specifični nizovi reči mogu biti jedinstveni za određenog autora, čime se značajno olakšava proces atribucije autorstva (Wright, 2017). Ova metoda je naročito korisna u slučajevima gde postoji sumnja na plagijat ili kada se analizira autentičnost tekstova kao što su anonimne poruke. Korpusni pristup omogućava lingvistima da sistematski analiziraju obrasce koji se često gube u tradicionalnim metodama analize.

Pored identifikacije autora, korpusi se koriste i za analizu leksičke frekvencije i stilskih varijacija unutar i između pravnih tekstova. Stanislav Gozdž-Roškovski je detaljno proučavao frazeologiju i varijacije u pravnim diskursima, ukazujući na to kako se određene fraze ponavljaju u specifičnim pravnim kontekstima, što je od suštinskog značaja za identifikaciju autora (Gozdž-Roszkowski, 2021, pp. 1521, 1526, 1529–1530). Korpusna analiza, takođe, omogućava otkrivanje evaluativnog jezika koji može biti ključan u pravnim postupcima.

2.3. Nedostatak korpusa na srpskom jeziku

Na srpskom jezičkom terenu, upotreba korpusnih alata u forenzičkoj lingvistici još uvek je u začetku. Iako su postignuti značajni pomaci u izgradnji srpskog korpusa, poput SrpKora (Utvić, 2014, str. 247 i dalje), specijalizovani forenzički korpusi još uvek ne postoje. Nedostatak ovakvih korpusa na srpskom jeziku predstavlja ozbiljan izazov i ograničava potencijal za istraživanja u oblasti FL, posebno u analizama autentičnosti tekstova i atribuciji autorstva.

Duško Vitas i saradnici prepoznaju značaj razvoja resursa za obradu srpskog jezika (Vitas et al., 2003), ali trenutni resursi nisu prilagođeni potrebama forenzičkih analiza. Na primer, složena morfološka struktura srpskog jezika, slobodan redosled reči u rečenici i prisustvo dva pisma (ćiriličnog i latiničnog) predstavljaju specifične izazove za obradu jezika (Ibid, p. 2). Ovi tehnički izazovi otežavaju razvoj alata koji bi mogli omogućiti efikasnu analizu tekstova u forenzičkoj praksi.

Nikola Dobrić (2012) ukazuje na to da su zapadnobalkanski jezici napravili značajne korake u razvoju korpusa, ali ipak postoji potreba za specifičnijim alatima koji bi omogućili detaljniju analizu pravnih i kriminalističkih tekstova. Odsustvo specijalizovanih forenzičkih korpusa na srpskom jeziku predstavlja prepreku u napretku FL, posebno u oblastima

kao što su identifikacija autora i provera autentičnosti dokumenata u pravnim sporovima.

Razvoj forenzičkog korpusa za srpski jezik bio bi od ključnog značaja za unapređenje pravnih istraživanja i analize tekstova. Takav korpus omogućio bi ne samo precizniju analizu spornih tekstova već i unapredio razumevanje jezika u pravnom kontekstu, što bi znatno olakšalo rad lingvista i pravnih stručnjaka.

3. Predlog metodologije

3.1. Kreiranje korpusa forenzičkih tekstova na srpskom jeziku

Kreiranje specifičnog korpusa forenzičkih tekstova na srpskom jeziku zahteva pažljivo planiranje i odabir relevantnih tekstova koji će služiti kao reprezentativan uzorak za analizu. Ovaj korpus obuhvata različite vrste tekstova koji su od ključne važnosti u pravnim i kriminalističkim kontekstima. Posebnu pažnju treba posvetiti uključivanju tekstova kao što su policijski izveštaji, pravni dokumenti, izjave, preteće poruke, oprostajna pisma, zahtevi za otkupninu i apeli. Među njima se izdvajaju izveštaji nadležnih organa kao jedan od najvažnijih tipova, jer sadrže detalje o događajima koji se istražuju u sudskim postupcima, uključujući izjave svedoka i osumnjičenih. Oni pružaju uvid u poseban stil poznat kao policijski registar koji se koristi u policijskim dokumentima, čineći ih neophodnim delom korpusa.

Pravni dokumenti, uključujući tužbe, žalbe, presude i druge sudske spise, predstavljaju primere pravnog jezika i diskursa ključnog za analizu jezika u pravnim procesima. Takođe, korpus bi obuhvatio preteće poruke i oprostajna pisma – tekstove koji se često koriste u forenzičkim istragama za utvrđivanje autentičnosti i identifikaciju autora. Ovi tekstovi omogućavaju uvid u neformalni jezik i stil koji se koristi u kriminalnim aktivnostima, što ih čini važnim za forenzičku analizu.

Korpus treba da bude dinamičan, s mogućnošću redovnog ažuriranja i dodavanja novih tekstova, čime će se obezbediti njegova aktuelnost i relevantnost u vremenu.

3.2. Kriterijumi selekcije tekstova

Svi tekstovi koji bi bili uključeni u korpus pažljivo će se birati prema unapred definisanim kriterijumima kako bi se obezbedila tačnost i relevantnost podataka. Autentičnost tekstova je od suštinskog značaja – u korpus treba uključiti samo one tekstove koji su prošli kroz zvanične pravne ili policijske postupke i koji su sastavni deo zaključenih sudskih

predmeta. Reprezentativnost tekstova je takođe ključna, pri čemu treba osigurati da uzorci pokrivaju različite vrste pravnih i kriminalnih slučajeva, kao i raznolike stilove pisanja.

Osim toga, posebna pažnja posvetiće se obezbeđivanju varijabilnosti u tekstovima, uključujući tekstove različitih registara, od formalnog pravnog jezika do neformalnog jezika anonimnih poruka. Konačno, svi podaci će biti anonimizovani kako bi se zaštitila privatnost aktera, a proces izgradnje korpusa biće u potpunosti usklađen s etičkim smernicama.

3.3. Tehnički aspekti i jezički alati

Korpus će biti kreiran korišćenjem naprednih softverskih alata za analizu jezika, koji omogućavaju detaljnu analizu različitih jezičkih elemenata. Jedan od ključnih aspekata biće analiza leksičke frekvencije, koja omogućava prepoznavanje obrazaca u učestalosti reči, identifikaciju ključnih termina i uvid u leksičku strukturu tekstova. Alati poput SrpNet, AntConc, Unitex, SketchEngine i NoSketchEngine, WordSmith Tools, koji se često koriste u analizi jezika, pružiće mogućnost za pretragu i obradu tekstova na srpskom jeziku (Utvić, 2014, str. 208–243, Vitas et al., 2012, pp. 23, 34–35). Takođe, alati koje razvija NLP grupa na Matematičkom fakultetu Univerziteta u Beogradu (up. Vitas et al., 2003, para. 4), pružiće dodatne mogućnosti za analizu srpskog jezika u kontekstu forenzičkih tekstova.

Sintaksička analiza biće sprovedena korišćenjem alata kao što je Unitex/GramLab, koji omogućava prepoznavanje sintaksičkih obrazaca i dodavanje oznaka (Vitas et al. 2003; Krstev & Vitas, 2005; Utvić, 2014), ali ne sprovodi potpunu analizu rečeničnih struktura. Ova analiza omogućići će istraživačima da identifikuju određene sintaksičke konstrukcije i stilske karakteristike specifične za forenzičke tekstove. Iako se alat koristi za leksičku i sintaksičku obradu, uključujući segmentaciju, normalizaciju i disambiguaciju teksta, nije direktno namenjen za stilometrijske tehnike poput n-gram modela. Sistem se primarno oslanja na konačne automate i transduktore, ali n-gram modeli i drugi statistički stilometrijski alati nisu deo njegovih osnovnih mogućnosti. U kontekstu srpskog jezika, prema Vitasu i saradnicima (2012), razvijeni su različiti jezički resursi i alati, ali se dodatno naglašava potreba za naprednim tehnologijama i alatima koji omogućavaju sveobuhvatnu sintaksičku i semantičku analizu za bolje razumevanje složenih jezičkih struktura, što uključuje i forenzičku primenu.

Alati kao što su Unitex/GramLab i TXM nude dodatne opcije za analizu srpskog jezika u kontekstu leksičke i sintaksičke obrade.

Unitex/GramLab je alat za obradu prirodnog jezika koji koristi konačne automate za prepoznavanje reči i sintaksičkih obrazaca. Može se primeniti na srpski jezik uz odgovarajuće prilagođavanje leksikona i gramatičkih resursa, ali njegova ograničenja uključuju nedostatak duboke sintaksičke analize i odsustvo stilometrijskih tehnika kao što su n-gram modeli. S druge strane, TXM je softver za tekstometrijsku analizu koji omogućava frekvencijsku analizu, klasterizaciju i vizuelizaciju leksičkih odnosa u tekstualnim korpusima, te zahteva anotirane korpuse u XML formatu. TXM podržava Unicode i TEI standarde, što ga čini pogodnim za analizu srpskih korpusa, uz mogućnost primene različitih statističkih metoda za analizu tekstualnih struktura. Iako oba alata nude korisne funkcije za osnovnu analizu, za dublje stilometrijske analize i primenu statističkih modela preporučuje se korišćenje dodatnih alata, kao što su Python biblioteke (NLTK, scikit-learn) ili Stylo za naprednu obradu i poređenje tekstova.

Za stilometrijsku analizu, koja omogućava identifikaciju specifičnih jezičkih obrazaca karakterističnih za određenog autora ili tip teksta, obično se koriste drugačiji alati specifično razvijeni za statističku obradu teksta, kao što su R, Python biblioteke (npr. NLTK ili scikit-learn) ili namenski stilometrijski alati poput Stylo. Istraživanje Dejvida Vrajta (Wright, 2017) pokazuje da n-gram modeli, posebno nizovi reči između dve i šest reči (n-gramovi), mogu biti korisni kao stilski markeri za identifikaciju autora u forenzičkim analizama jer pružaju mogućnost poređenja specifičnih jezičkih obrazaca među tekstovima u različitim korpusima.

Srpski jezik se suočava sa specifičnim izazovima kada je reč o dostupnosti jezičkih alata za forenzičku jezičku analizu. Iako postoje nacionalni korpusi kao što je SrpKor, nedostatak specijalizovanih forenzičkih alata kao, na primer, onih koji bi omogućili duboko parsiranje a koji još uvek nisu razvijeni za srpski jezik (Vitas et al. 2012, p. 34), i dalje ograničava mogućnosti za dubinsku analizu tekstova u ovoj oblasti. Razvoj dodatnih softverskih alata prilagođenih potrebama FL na srpskom jeziku, kao i implementacija stilometrijskih tehnika u okviru postojećih sistema kao što je Unitex/GramLab i NooJ, biće ključni za unapređenje ove oblasti i poboljšanje efikasnosti forenzičkih analiza.

4. Analiza i diskusija

4.1. Mogućnosti analize forenzičkih tekstova

Korpus forenzičkih tekstova na srpskom jeziku nudi brojne mogućnosti za analizu autentičnosti, identifikaciju autora i pružanje jezičkih dokaza u pravnim postupcima. Analiza autentičnosti omogućava utvrđivanje verodostojnosti teksta, a u korpusu se mogu identifikovati stilski markeri specifični za određene vrste tekstova ili autorov jezički profil. Korišćenjem leksičke i stilometrijske analize, forenzički lingvisti mogu uporediti tekstove koji pripadaju različitim pravnim žanrovima i utvrditi da li tekst sadrži odstupanja koja bi ukazivala na manipulaciju ili falsifikovanje.

U domenu identifikacije autora, korpus bi omogućio analizu idiolekatskih elemenata kroz upotrebu n-gramova i ponavljajućih fraza, čime se može identifikovati specifičan jezički obrazac karakterističan za pojedinca ili grupu. Na primer, istraživanja su pokazala da nizovi reči (n-gramovi) omogućavaju identifikaciju specifičnih jezičkih obrazaca koji se koriste za razlikovanje autentičnih tekstova od onih s nepoznatim autorom (Wright, 2017). Upotrebom ovih analiza, moguće je povezati sporne tekstove s potencijalnim autorima, čime se forenzički korpus pozicionira kao ključan alat u krivičnim istragama.

4.2. Praktični primeri i mogućnosti identifikacije autorstva

Praktična primena korpusa u identifikaciji autorstva može se prikazati kroz nekoliko konkretnih primera. Kada se ispituje autentičnost ovakvih tekstova, analiza stilskih markera — kao što su specifični nizovi reči, ton i struktura — može otkriti podudarnosti s jezičkim obrascima poznatih autora. Na primer, u slučaju poruke u kojoj je bila zahtevana uplata u zamenu za kompromitujući dokument s dokazima, korišćena je konstrukcija s leksičkim izborom koji bi mogao ukazivati na prisustvo idiolekta: *Stavite novac u kovertu, 1.000 € u dunavskom parku iza izviđačkog kampa, cd vam je zaboden između zida tu osavite kovertu a uzmite cd mozete doci odmah cd je vec tamo*. U ovakvom slučaju važno je utvrditi da li je reč o idiolektu, odnosno kolika je frekvencija ovakvog leksičko-gramatičkog spoja. Drugim rečima, pretragom korpusa s ciljem poređenja jezika analiziranih tekstova sa referentnim primerima, mogu se identifikovati obrasci koji nisu uobičajeni ili odstupaju od norme što bi pomoglo u donošenju preciznih zaključaka o potencijalnom autoru.

U ovakvim primerima, važno je utvrditi i da li pretnja ima realnu osnovu, odnosno srž, koja ukazuje na stvarnu opasnost ili je pretnja

usmerena na zastrašivanje bez konkretne namere za izvršenje. Praktična primena korpusa u identifikaciji autorstva i proceni ozbiljnosti pretnji može biti ključna za utvrđivanje da li poruka sadrži realnu pretnju koja zahteva dalju istragu. Na osnovu klasifikacije pretnji prema stepenu rizika, FBI je razvio smernice koje omogućavaju procenu ozbiljnosti pretnji putem analize specifičnih lingvističkih elemenata. Na primer, pretnje s visokim stepenom rizika obuhvataju direktne, uverljive izjave u vezi sa žrtvom i često uključuju detalje o oružju, mestu i vremenu napada, što signalizira da je preduzeta priprema za ostvarivanje pretnje (Nikolić Novaković, 2017, str. 104–105), dok pretnje kao što je *J***m ti mater, budalo matora, j****u ti majku, ubicu te, sve cu vas pobiti, za sve si ti kriva ljubice j****u ti majku, ubicu te, sredicu Vas Da li bi zelela da zaradis 20 eura, dobro si parce, gde ti je majka, jel otisla da se prodaje u belo roblje i tebe odvucla, sve cu vas pobiti*, koje uključuju uvredljive i vulgarne izraze, ali bez jasnih detalja o planiranom napadu, najčešće ukazuju na manji rizik jer nemaju srž. Takve pretnje se uglavnom karakterišu afektivnim izrazima besa i namerom da izazovu strah bez čvrste osnove za realizaciju. Upotrebom stilometrijskih metoda i procene rizika u analizi pretećih tekstova, korpus bi mogao omogućiti istražiteljima da razlikuju pretnje koje imaju visok stepen verodostojnosti od onih koje služe samo za zastrašivanje, čime se doprinosi preciznosti i efikasnosti u kriminalističkim istragama.

U sličnom kontekstu, analiza oproštajnih pisama može otkriti jedinstvene stilske elemente koji bi mogli ukazivati na lažno autorstvo. Prava oproštajna pisma obično imaju karakteristike kao što su izrazi ljubavi prema porodici, direktivna uputstva najbližima i odsustvo populističkih stavova poput iskazivanja sopstvene slabosti ili kukavičluka. Ove osobine razlikuju autentične poruke od fingiranih oproštajnih pisama, koja mogu sadržavati te elemente da bi izazvala emocionalni odgovor kod čitaoca ili lažno predstavila autorov stav i osećanja. Korisnost postojanja korpusa ogleda se u mogućnosti poređenja jezika analiziranih tekstova s referentnim korpusom kako bi se identifikovali obrasci koji odstupaju od norme, što doprinosi preciznosti analize i donošenju mišljenja o autentičnosti teksta. U jednom primeru iz stvarne prakse, samoubica je ostavio sledeće pismo: *Neno moja, ne mogu više da izdržim. Plačem za svim što sam izgubio. Naše anđele i tebe. Za sve sam ja kriv. Čuvaj naše devojčice. Sve vas volim puno*. Ovakva frazeologija, koja sadrži izraze sopstvene krivice, tuge i brige za porodicu, oslikava autentičan emotivni ton i privrženost najbližima, što je indikator autentičnosti oproštajnog pisma.

4.3. Uloga sintaksičke analize u identifikaciji autora

Prema Kerol Časki, sintaksička analiza može značajno doprineti identifikaciji autora u forenzičkim istraživanjima, posebno kada se radi o kratkim, fragmentiranim izjavama. Ona predlaže stilometrijske metode zasnovane na kvantitativnoj analizi sintaksičkih struktura kao što su dužina reči, frekvencija rečeničnih struktura i raspored sintaksičkih markera, pri čemu se postiže visoka tačnost u razlikovanju autora, čak i kada je količina podataka ograničena (Časki, 2001, 2005). U okviru ovih metoda nekada se koristio softver ALIAS, koji je omogućavao lematizaciju, proračunavanje frekvencije reči, kategorizaciju interpunkcije i sintaksičku analizu. Program je bio zasnovan na sintaksičkom pristupu autorskoj atribuciji i obuhvatao je automatsku segmentaciju teksta na rečenice, označavanje gramatičkih kategorija (Part-Of-Speech tagging), kategorizaciju interpunkcijskih znakova i identifikaciju sintaksičkih struktura. Međutim, ALIAS se danas ne koristi budući da su u međuvremenu razvijeni savremeniji alati zasnovani na generativnoj veštačkoj inteligenciji, posebno veliki jezički modeli, koji se pokazali kao izuzetno efikasni u stilometrijskim i forenzičkim analizama različitih tekstualnih zadataka (Michelet & Breitinger, 2024). Ovi modeli omogućavaju dublju obradu teksta i preciznije razlikovanje autora, naročito u tekstovima sa nepoznatim autorstvom.

Istraživanja K. Časki ukazuju na značaj prepoznavanja specifičnih jezičkih struktura kao stilometrijskih markera koji otkrivaju jedinstvene jezičke obrasce karakteristične za pojedince, što je od izuzetne važnosti za formiranje i primenu forenzičkog korpusa na srpskom jeziku. Takvi markeri mogu biti presudni u preciznoj identifikaciji autora i u proceni legitimnosti spornih tekstova. Uvođenje ovih modela u analizu tekstova na srpskom predstavljalo bi značajan iskorak u razvoju forenzičke lingvistike.

Metodološki pristup Časki zasniva se na sintaksičkoj klasifikaciji interpunkcijskih znakova kao markera autorstva i naglašava značaj sofisticiranih jezičkih alata koji mogu da razlikuju suptilne varijacije unutar tekstova, uključujući raspored sintaksičkih granica sintagmi i rečenica (Časki, 2005, pp. 5–7). Kombinacijom tradicionalnih stilometrijskih metoda i naprednih sintaksičkih analiza postiže se veća tačnost i pouzdanost u identifikaciji autora u digitalnim dokazima (Časki, 2001, pp. 2–3, 7), što bi činilo srpski forenzički korpus vrednim resursom u istragama u kojima su identitet autora i autentičnost teksta ključni za pravne postupke.

U tom kontekstu, posebno je važan pojam markiranosti, koji u lingvistici označava asimetričnu binarnu opoziciju — jedan element (nemarkirani) predstavlja neutralnu ili češću formu, dok je drugi

(markirani) ređi, formalno obeležen ili stilski naglašen. U sintaksi, nemarkirane strukture su uobičajene i lakše za obradu, dok markirane mogu biti složenije i ređe. Upravo zahvaljujući tim razlikama, sintaksička markiranost može omogućiti identifikaciju autorstva unutar istog autora, kao i razlikovanje između različitih autora (Chaski, 1997).

5. Potencijalni izazovi, bezbednosni i etički aspekti

U radu s tekstovima kao što su policijski izveštaji, pretnje, izjave svedoka i drugi pravno relevantni dokumenti, nezaobilazna su pitanja zaštite podataka. Iako su pravni aspekti privatnosti regulisani važećim zakonodavstvom, uključujući i Opštu uredbu o zaštiti podataka (GDPR),³ u savremenom digitalnom okruženju sve veću pretnju predstavljaju bezbednosni rizici, posebno hakerski napadi. Izveštaji vladinih institucija i pravni dokumenti često su meta sajber napada, čime se ugrožava integritet i poverljivost forenzičkog materijala (Mayer, 2017). Stoga je pri kreiranju i korišćenju digitalnog korpusa neophodna primena visokih standarda zaštite, uključujući enkripciju, kontrolu pristupa, pseudonimizaciju osetljivih podataka i redovno ažuriranje bezbednosnih protokola. Pored tehničkih mera, važno je istaći etičku odgovornost istraživača u sprečavanju moguće zloupotrebe podataka. Forenzičkolingvistička analiza, naročito u pravnim postupcima, može imati direktne posledice po pojedince. Zbog toga je neophodno osigurati transparentnost metodologije, informisanost relevantnih strana i minimizovanje rizika od pogrešne interpretacije ili instrumentalizacije jezičkih dokaza.

6. Zaključak

Ovaj rad doprinosi razvoju FL u Srbiji, posebno u kontekstu kreiranja i primene digitalnog specijalizovanog korpusa forenzičkih tekstova na srpskom jeziku. Istaknute su mogućnosti koje ovakav korpus pruža u analizama autentičnosti teksta, identifikaciji autora i rešavanju pravnih sporova upotrebom jezičkih dokaza. Korpus bi obuhvatio raznovrsne forenzičke tekstove, uključujući policijske izveštaje, pretnje i oproštajna pisma, čime bi se omogućilo dublje razumevanje različitih jezičkih obrazaca unutar kriminalističko-pravnih diskursa.

Dodatna vrednost rada ogleda se u metodologiji koja bi se razvila za formiranje korpusa, kao i u uspostavljanju kriterijuma za selekciju

³ General Data Protection Regulation (GDPR) je pravni okvir Evropske unije koji reguliše prikupljanje, obradu i čuvanje ličnih podataka (European Union, 2016). U kontekstu forenzičke lingvistike, GDPR zahteva strogu kontrolu pristupa i obrade tekstova koji sadrže podatke o ličnosti, posebno u slučajevima koji uključuju osetljive pravne ili policijske informacije.

relevantnih i autentičnih tekstova. Takođe, u radu je ukazano na važnost prilagođavanja softverskih alata specifičnostima srpskog jezika koji se suočava sa složenom morfologijom i slobodnim redosledom reči. Ovaj pristup ne samo da omogućava analizu specifičnih diskursa već doprinosi i razvoju alata za obradu srpskog jezika, što ima široko primenljiv potencijal.

Na osnovu dosadašnjih nalaza, preporučuje se nastavak rada na stvaranju sveobuhvatnog forenzičkog korpusa na srpskom jeziku, uz proširenje na dodatne vrste tekstova poput zahteva za otkupninu i krivičnih prijava. Dalji razvoj ovakvog korpusa zahteva interdisciplinarnu saradnju lingvista, pravnika i IT stručnjaka kako bi se osigurala precizna analiza i poštovanje zakonskih regulativa o zaštiti podataka.

Preporučuje se da buduća istraživanja budu usmerena na razvoj i primenu naprednih softverskih rešenja koja bi mogla automatizovati procese računanja leksičke frekvencije, sintaksičke analize i stilometrijskih modela specifičnih za srpski jezik. Korišćenje savremenih NLP tehnika, kao što su n-gram modeli i prepoznavanje jedinstvenih jezičkih obrazaca, dodatno bi unapredilo tačnost identifikacije autora i verodostojnosti teksta u pravnim kontekstima.

Na kraju, neophodna je kontinuirana edukacija stručnjaka u vezi s etičkim smernicama i zaštitom privatnosti prilikom obrade forenzičkih tekstova. Uspostavljanje korpusa predstavlja osnovu za dalji razvoj forenzičke lingvistike u Srbiji, a njegova implementacija bi značajno unapredila kvalitet i pouzdanost pravnih analiza, doprinoseći globalnoj mreži forenzičkih istraživanja i pravnih resursa.

Literatura

- [1] Blackwell, S. (2009). Why forensic linguistics needs corpus linguistics. *Comparative Legilinguistics*, 1, 5–19. <https://doi.org/10.14746/cl.2009.01.01>
- [2] Chaski, C. (1997). Who wrote it? Steps toward a science of authorship identification. *National Institute of Justice Journal*, 233(233), 15-22.
- [3] Chaski, C. E. (2001). Empirical evaluations of language-based author identification techniques. *Forensic linguistics*, 8, 1–65.
- [4] Chaski, C. E. (2005). Who's at the keyboard? Authorship attribution in digital evidence investigations. *International journal of digital evidence*, 4(1), 1–13.
- [5] Coulthard, Malcolm (2005). Some forensic applications of forensic linguistics. *Revista Veredas de Estudos Linguisticos*, 9, 9–28.
- [6] Coulthard, M. (2013). On the use of corpora in the analysis of forensic texts. *International Journal of Speech, Language and the Law*, 1(1), 27–43.

- [7] Dobrić, N. (2012). Savremeni jezički korpusi na zapadnom Balkanu–istorijat, trenutno stanje i budućnost. *Slavistična revija*, 60(4), 677–692. https://srl.si/ojs/srl/article/view/COBISS_ID-51417186
- [8] European Union. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation). *Official Journal of the European Union*, L 119, 1–88. <https://eur-lex.europa.eu/eli/reg/2016/679/oj>
- [9] Goźdz-Roszkowski, S. (2021). Corpus linguistics in legal discourse. *International Journal for the Semiotics of Law-Revue internationale de Sémiotique juridique*, 34(5), 1515–1540. <https://doi.org/10.1007/s11196-021-09860-8>
- [10] Krstev, C. & Vitas, D. [2005]. Corpus and Lexicon – Mutual Incompleteness. In P. Danielsson & M. Wagenmakers (eds), *Proceedings of the Corpus Linguistics Conference*, Vol. 4, Birmingham. Retrieved May 6, 2025 from https://www.researchgate.net/profile/Cvetana-Krstev/publication/266882123_Corpus_and_Lexicon_-_Mutual_Incompleteness/links/544792880cf22b3c14e0ed6b/Corpus-and-Lexicon-Mutual-Incompleteness.pdf
- [11] Michelet, G., & Bretinger, F. (2024). ChatGPT, Llama, can you write my report? An experiment on assisted digital forensics reports written using (local) large language models. *Forensic Science International: Digital Investigation*, 48, 301683. Retrieved May 6, 2025 from <https://www.sciencedirect.com/science/article/pii/S2666281723002020>
- [12] Mayer, J. (2017). Government hacking. *Yale LJ*, 127, 570.
- [13] Nikolić-Novaković, L. (2017). *Forenzička lingvistika. Primena metoda forenzičke lingvistike u kriminalističkim istragama pretnje*. Banjaluka: Evropski defendologija centar za naučna, politička, ekonomska, socijalna, bezbjednosna, sociološka i kriminološka istraživanja.
- [14] Olsson, J. (2010). *Forenzička lingvistika*. Zagreb: Globus.
- [15] Svartvik, J. (1968). *The Evans statements: A case for forensic linguistics*. Göteborg: University of Göteborg. Retrieved December 20, 2020 from <https://www.thetext.co.uk/Evans%20Statements%20Part%202.pdf>
- [16] Utvić, M. V. (2014). *Izgradnja referentnog korpusa savremenog srpskog jezika* (Publication No. 30969143) [Doktorska disertacija, University of Belgrade]. ProQuest Dissertations & Theses.
- [17] Vitas, D., Krstev, C., Obradovic, I., Popovic, L., & Pavlovic-Lazetic, G. (2003). An overview of resources and basic tools for processing of Serbian written texts. In *Proc. of the Workshop on Balkan Language Resources, 1st Balkan Conference in Informatics*.
- [18] Vitas, D., Popović, L., Krstev, C., Obradović, I., Pavlović-Lažetić, G., & Stanojević, M. (2012). *The Serbian language in the digital age*. META-NET

White Paper Series, G. Rehm, H. Uszkoreit (eds.).

<http://dr.rgf.bg.ac.rs/s/repo/item/0000764>

- [19] Vorkapić, D., Tomašević, A., Mladenović, M., Stanković, R., & Vulović, N. (2017). Digital Library From A Domain Of Criminalistics As A Foundation For A Forensic Text Analysis. In *International Scientific Conference “Archibald Reiss Days” Thematic Conference Proceedings Of International Significance*, Belgrade, 7–9 November 2017.
- [20] Wright, D. (2017). Using word n-grams to identify authors and idiolects: A corpus approach to a forensic linguistic problem. *International journal of corpus linguistics*, 22(2), 212–241.

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

Jelena Redli

Summary

Forensic linguistics (FL) is an interdisciplinary field that combines linguistic, legal, and digital technologies to analyze language use in legal and criminal contexts. Its primary applications include author identification, plagiarism detection, and criminal investigations, relying heavily on specialized linguistic corpora. However, forensic linguistic research on the Serbian language remains underdeveloped due to the absence of dedicated forensic text corpora, limiting the applicability of advanced analytical methods.

This paper investigates the potential for developing and applying a forensic text corpus for Serbian, highlighting its role in text authenticity verification, authorship attribution, and the resolution of legal disputes through linguistic evidence. The proposed methodology involves compiling and analyzing various types of forensic texts, including police reports, legal documents, threatening messages and suicide notes. The study also explores the integration of advanced computational linguistic tools for automated text analysis, including lexical frequency profiling, syntactic parsing, stylistic marker detection, and forensic stylometry techniques.

A key challenge in constructing a Serbian forensic corpus is addressing the linguistic complexity of the language, including its rich morphological structure, free word order, and dual script usage (Cyrillic

and Latin). Additionally, the research identifies major technical, legal, and ethical barriers in corpus development, particularly regarding data privacy, anonymization, and the ethical use of sensitive legal texts. The paper underscores the necessity of interdisciplinary collaboration between linguists, legal experts, law enforcement agencies, and IT professionals to ensure the corpus is both methodologically rigorous and legally compliant.

Beyond theoretical contributions, the paper presents potential real-world applications of a Serbian forensic corpus, such as its use in criminal investigations, fraud detection, and threat assessment. By analyzing forensic texts with computational methods, law enforcement and judicial authorities could improve the accuracy of authorship identification and linguistic profiling. The study also raises critical questions for future research, including:

How can computational forensic linguistic tools be adapted to Serbian's linguistic structure?

What legal frameworks have to be established to protect privacy while enabling forensic text analysis?

How can forensic corpora be continuously updated to reflect emerging linguistic trends in digital communication?

The findings suggest that establishing a specialized forensic corpus for Serbian would significantly enhance forensic linguistic research and its application in legal practice. Furthermore, the creation of such a resource would align Serbian forensic linguistics with global trends, facilitating cross-linguistic comparisons and the advancement of digital forensic methodologies. The paper concludes with a call for interdisciplinary efforts to bridge the gap between linguistic research, law enforcement needs, and technological advancements in forensic text analysis.

Keywords: forensic linguistics, Serbian language, legal language, digital corpus, linguistic tools, forensic text analysis, authorship identification.

Sonority Based Syllabic Hyphenation of Macedonian and Serbian

Scientific paper

DOI: 10.18485/judig.2025.1.ch5

Katerina Zdravkova,  0000-0002-9674-3081

*Jana Kuzmanova*¹  0009-0009-5015-4183

Abstract

Hyphenation is a technique of joining or separating syllables of a word by adding a hyphen. This is typically achieved by implementing the syllabic segmentation. We have attempted to achieve the same task based on the sonority of phonemes, considering that sonority within a syllable increases to the syllable nucleus and decreases at syllable boundaries. For Macedonian, this principle was extended by an additional rule stating that a syllable begins when a monotonically non-decreasing sonority ends. In Serbian, monotonic non-decreasing starts at the beginning of a new syllable. By defining appropriate sonority weights for both languages and defining a very simple splitting strategy, the accuracy of the syllabic hyphenation has surpassed the rule-based approach. It reached 97.59% for Macedonian and 98.68% for Serbian. We intend to further improve it by taking into account PoS tags for Macedonian and to fine-tune the sonority weights for Serbian, hoping to achieve an accuracy that exceeds 99%.

Keywords: Macedonian, Serbian, syllabic hyphenation, phoneme sonority, ChatGPT.

1. Introduction

Hyphenation is a technique of joining or separating the syllables of words that improves their legibility, readability, aesthetics and visual balance of printed or displayed texts [1]. The separation is done at the end

¹ Ss Cyril and Methodius University in Skopje, Faculty of Computer Science and Engineering, Skopje, {katerina.zdravkova; jana.kuzmanova}@finki.ukim.mk

of a line, using a hyphen to indicate where the break occurs. Since this process combines the concepts of both syllabic and hyphenation, the term for this process can be generalized as syllabic hyphenation [2]. It follows the natural syllable boundaries within a word in a way that respects their linguistic structure [3]. In the languages with a highly predictable syllabic structure, hyphenation provides the most natural and linguistically logical way to break words. In the languages with more complicated compound structures or where typography is heavily involved, non-syllabic hyphenation might be used in specific contexts [4].

The hyphenation of Macedonian language was embedded into Korektor-M, the first software package for spell checking and hyphenation of Macedonian texts. Created by Dragan Mihajlov and Dejan Gjorgjevikj more than 25 years ago, Korektor-M was a favourite tool of many newspapers that used it for automatic proofreading. With the development of Microsoft Office, it became useless. However, although new office suites offer spell checking option, word hyphenation of Macedonian language is not yet enabled.

The first attempt to create a new system for Macedonian hyphenation was done in 2023 [5]. It consisted of two mutually separate parts supporting syllabification and morpheme segmentation [5]. The segmentation was entirely based on rules that are quite general and ambiguous (<https://makedonskijazik.mk/>). The accuracy of the approach was below our expectations. Therefore, an alternative approach based on phoneme sonority was proposed [6], which was presented on JuDig conference (<https://judig.jerteh.rs/>). It is explained in more detail in this paper.

The hyphenation of Serbian language has a much longer tradition. It started with the research done by Krstev [7] and was then extended as part of word length counting [8]. The resources used for measuring the length of syllables in Serbian [9] were the bases for evaluating the proposed hyphenation approach. Similarly to Macedonian, the rule based approach was rather complex [10]. Exceeding the 98% accuracy obtained during this study became our main challenge.

2. Sonority Based Syllabic Hyphenation

Phonetically, syllables are sequences of sounds containing one peak of prominence [5]. Phonologically they are units of stress placement. According to the Sound Sequencing Principle, sonority within a syllable rises to the nucleus of the syllable and then falls in sonority [11].

The sonority of Macedonian phonemes depends on their basic classification: vowels, sonorants, voiced and voiceless consonants (Table 1). The sonority of Serbian phonemes is more extensive and includes three additional classes: plosives, fricatives and affricates, in each of which phonemes can be voiced or voiceless (Table 2).

Table 1: Sonority of Macedonian phonemes

Phoneme class	Phonemes	Sonority weight
Vowels	<i>a e u o y</i>	12
Sonorant <i>r</i>	<i>p</i>	6
Sonorants	<i>j л љ м н њ</i>	4
Voiced consonants	<i>б в г д р ж з с ш</i>	2
Voiceless consonants	<i>к п ц т к' х џ ч џ'</i>	1

Table 2: Sonority weights of Serbian phonemes

Phoneme class	Phonemes	Sonority weight
Vowels	<i>a e u o y</i>	12
Sonorant <i>r</i>	<i>p</i>	8
Sonorants <i>l, m</i> and <i>n</i>	<i>л м н</i>	6
Other sonorants	<i>в ј љ њ</i>	5
Plosive voiced	<i>б г д</i>	4
Plosive voiceless	<i>к п т</i>	3
Fricative voiced	<i>з ж</i>	3
Fricative voiceless	<i>с х ф ш</i>	2
Affricates voiced	<i>ђ ћ њ</i>	2
Affricates voiceless	<i>џ ч</i>	1
Special signs	FC S	0

The syllable nuclei in both languages are the five vowels. Their sonority weight is set to 12. In Macedonian, a nucleus can be the sonorant *p* (Latin transcription: *r*) appearing within a consonant group (*крст*, *врста*, *пр-веш-ство*) or at the end of the word (*ма-са-кр*). In Serbian language, apart from the sonorant *p* (*мерп*, *црв*, *тр-ка*), the sonorants *л*

and *h* can also become syllable nuclei (for example, *би-ци-кл*, *де-ба-кл*, *Вл-та-ва*, *Њу-тн*).

Macedonian language has a special sound ‘, which appear in words such as: ‘*рж*, ‘*рџа*, and ‘*рџем*, mainly at the beginning of the word and succeeded by the sonorant *p*.

Vowel groups (for example, *ay* in *а-у-по-па*, *ea* in *и-де-ал*, and *uo* in *а-ви-он*) in both languages are separated by a fictive consonant FC with a sonority weight of 0. Word delimiters are assigned as S, with the sonority weight of 0.

The syllabic roles of the phonemes are determined by calculating the triplet difference:

$$TD(p_i) = w(p_i) - w(p_{i-1}) - w(p_{i+1}), i = 1, \dots, n \quad (1)$$

The triplet difference of the nuclei has always a positive value, while the consonants from the onset and the coda are always negative (Table 3 and Table 4). The only exclusion to this rule in Macedonian are the words starting with a voiced consonant succeeded by a voiceless consonant, such as *вчера* and *вчудовиден*. The problem was resolved by adding a restriction that a single voiced consonant cannot form a separate syllable.

In Serbian language, the high sonority weight of the sonorant *p*, when it is between two vowels, reduces the triplet difference of the vowel that immediately follows the fictive consonant, as in *а-у-по-па*, making it negative. This relatively rare inconsistency was solved in the second iteration of the syllabic hyphenation of Serbian, in which the weight of the non-syllabic *p* was reduced to 6.

The determination of syllable boundaries depends on a monotonicity of sonority weights. In Macedonian, a new syllable starts when monotonic non-decreasing ends. According to this constraint, whenever the sonority of two adjacent Macedonian consonants is identical, in that case the second one belongs to a new syllable (Table 3). In Serbian, a new syllable begins at a phoneme where the sonority weight series begins to monotonically increase (Table 4).

Table 3: Syllabic hyphenation of the word *идеално* in Macedonian

	S	И	Д	Е	FC	А	Л	Н	О	S
phonemme										
sonority	0	12	2	12	0	12	4	4	12	0
triplet difference		10	-22	10	-24	8	-12	-12	8	

Table 4: Syllabic hyphenation of the word *идеално* in Serbian

	S	И	Д	Е	FC	А	Ј	Н	О	S
phonemme										
sonority	0	12	4	12	0	12	6	6	12	0
triplet difference		8	-20	8	-24	6	-12	-12	6	

3. Development of Syllabic Hyphenation of Macedonian and Serbian

The syllabic hyphenation was developed in Python. The algorithm used in all experiments consists of the same general steps with adjustments for each language. The steps in the base algorithm are the following:

1. Prepare the word – mark the beginning, end and put a marker between each two consecutive vowels.
2. Find the nuclei of the word.
3. Iterate over the characters in the word and form the syllables.

The steps to find the nuclei of the word are similar for both languages. First, the triplet difference is calculated for each group of three letters in the word, and then the nuclei are detected based on this calculation. Every vowel that has a nonnegative triplet difference with its preceding and following letter is detected as a nucleus, as well as the sonorant *p* surrounded by two consonants. For Serbian, the same rules are used, but additionally, the phonemes *л* and *н* can be nuclei if they have a positive triplet difference with the preceding and following phonemes.

The main difference between the experiments is in the way syllables are formed. In all cases, the formation of the syllable begins by adding phonemes until a nucleus is detected. In the baseline algorithm, this is followed by adding phonemes as long as the sonority of the current phoneme is higher than that of the following. If the sonority is not greater than that of the following phoneme, and the following phoneme is a consonant, then the first of the two phonemes is also added to the current syllable. This finalizes the formation of the syllable, and the algorithm continues by considering the next nucleus and again adding phonemes, starting one after the last added phoneme until this next nucleus is found.

The updated algorithm for Macedonian adds an additional rule to handle suffixes that shouldn't be split. These suffixes don't always appear at the end of a word – they can also appear in the middle, for example in declined forms of the word. This is done by simply looking at the phonemes surrounding the current phoneme. These are detected after the step of

splitting two consonants with the same sonority. The check happens when the first phoneme of the wanted suffix is added to the current syllable. If the following 2, 3 or 4 phonemes, along with the current, form one of the suffixes that should be kept together (*cmbo*, *cmвен*, *cku* and variations), then the current syllable is shortened and its last added phoneme is removed. In the next iteration, the first phoneme of the group is added to the syllable along with the other consonants that would usually be split, since they all come before the nucleus of the following syllable.

The algorithm for Serbian makes only one change to this algorithm. Here, the step where two consonants with the same sonority are split is skipped, and as soon as a consonant with a sonority that's not greater than that of the following is found, it directly checks if the word contains the groups that should be kept together, which are the same as for Macedonian.

4. Evaluation of Sonority Based Syllabic Hyphenation

The syllabic hyphenation of both languages started with the baseline model, which was our benchmark according to which we modified the approach. In Macedonian, it consisted of revoking the segmentation of the suffixes *cku*, *cmbo* and *cmвен* and their inflections. Namely, according to the orthographic rules they should remain within one syllable, which contradicts our constraint of splitting the adjacent phonemes with identical sonority weights into two syllables. In Serbian, the adjustments were done by decreasing the sonority weight of non-syllabic *p*.

The Macedonian language assessment sample consisted of 1310 words within average 3.09 syllables. The accuracy of the baseline algorithm was 89.26%, mainly due to the frequent occurrence of words with the suffixes *cku*, *cmbo* and *cmвен* and their inflections. By adjusting this, as explained in the previous section, the accuracy reached 96.63%. However, it affected the syllabification of the nouns: *гус-ки*, *мас-ки*, *прас-ки*, in which *cku* is not a morpheme. The only solution for effectively solving the problem of separable or non-separable suffix *ski* is the presence of the PoS tag word. This annotation overcomes the hyphenation problem.

An additional problem was caused by the ambiguous hyphenation of words with at least two adjacent consonants, such as the words: *автократ*, *Белград*, *декември* and *софтвер*. According to our approach, they were hyphenated as: *ав-ток-рат*, *Белг-рад*, *де-кемв-ри* and *соф-твер*, instead of the hyphenation suggested by the linguists: *ав-то-крат*, *Бел-град*, *де-кем-ври* and *софт-вер*. None of these automatic segmentations is incorrect as far as syllables are concerned. Namely, *крат*, *град* and *софт* are morphemes and they can be subject of further

segmentation, depending on the context. Still, when it comes to hyphenation, it is better to leave them as a whole. Introduction of morphemes will be one way to improve the system.

We were curious to find out how ChatGPT copes with the same task. Without any previous training, ChatGPT performed much below our expectations. Namely, many of the words were transliterated into Latin script (for example, Au-gust instead of ав-густ and cada-stre for ка-та-стра), presenting non-existent words. Many words had a wrong spelling, for example *абонент* was hyphenated as *а-бо-нет* instead of *а-бо-нент*, *авангарда* as *а-ва-ган-да* instead of *а-ван-гар-да* or *атрибум* with *а-му-бум* instead of *ат-ри-бум*. We then instructed ChatGPT with the rules for syllabic hyphenations and with the explanation to take care of the spelling and the Cyrillic script. Even with these instructions, the accuracy of hyphenation reached a modest 61.71%. Interestingly, after our explanations of what went wrong, this large language model gave examples confirming that it agrees with our suggestions, but repeated the same mistakes again, sometimes more than three times in a row.

The Serbian sample consisted of 3020 manually syllabified Serbian words. We first started with the same approach used for the Macedonian language, by simply modifying the sonority weights of the phonemes. After obtaining an accuracy of 75.34%, the constraint for the segmentation when the monotonic non-decreasing ends was adjusted with the constraint that the new syllable starts when the monotonic increase starts. With this adjustment, the accuracy of the baseline algorithm reached 97.59%, which was quite satisfactory. Most of the mistakes were related to words with non-sonoric *p*: *ар-ми-ја*, *мор-нар* and *у-пор-но* instead of *а-рми-ја*, *мо-рнар* and *у-по-рно*. Additionally, the inseparable Macedonian suffix *ски* has caused problems in the Serbian language as well, such as in the words: *гра-дског*, *љу-дски* and *сов-је-тска*. Apart from the fact that it was not segmented, it also collected the consonant that precedes it into a syllable. Several words containing the phoneme *к* (*лак-ше* instead of *ла-кше*, or *пот-сме-хом* instead of *по-тсме-хом*) were also not correctly syllabified. The last problem was the consonant group *num*, which should remain as a whole: *о-нуме*, *са-о-нуми*, and *у-о-нуме*.

By modifying the sonority of non-syllabic *p*, i.e. *p* which is not between two consonants from 8 to 6, the accuracy reached 98.68%, exceeding the rule-based syllabification accuracy [10]. Fine tuning of the sonority weights will be our next attempt to further improve the accuracy.

Similarly to Macedonian, we again tested ChatGPT. The accuracy of Serbian hyphenation was slightly better, reaching an acceptable 79.70% accuracy. Transliteration existed in a more subtle way: *ба-jo-net*, and *бы-dem*.

Spelling mistakes also occurred: *ве-чи-ном* and *бу-даш* instead of *ве-ћи-ном* and *бу-деи*. Consonants with the same sonority were separated: *ак-тив*, *бол-ни-це* and *бе-леж-ник* instead of *а-ктив*, *бо-лни-це* and *бе-ле-жник*.

When we informed ChatGPT that regardless of the training it had successfully completed, it had poorly realized the hyphenation of these two Slavic languages, the chatbot suggested that we try to repeat the same experiment with the professional version of Get Pro, the price of which is 200 US\$ per month. For now, we decided to leave this offer, especially because our system is still in the development phase.

5. Conclusions and Further Work

Novacula Occami or Occam's razor [12] teaches us that the simplest solution is often the best one. In this study, we have seen that the most straightforward approach to syllabic hyphenation for Macedonian and Serbian, based on the sonority of the languages, is both practical and efficient. By prioritizing simplicity, we have achieved a robust model for hyphenation that respects the core principles of syllabic structure without unnecessary complexity.

The main reason contributing to the high accuracy of syllabic hyphenation is the phonemic orthography of both languages where one phoneme is represented by one grapheme in the Cyrillic script. The languages are closely related but have different syllabification patterns. In Macedonian, hyphenation is based more on the phonemic sonority, where sounds of the same sonority are separated, while in Serbian, the syllable boundary tends to preserve a stronger connection between phonemes, regardless of their phonetic features.

The approach we proposed is extremely simple and at the same time, very efficient. We intend to further improve it by taking into account the PoS tags for the Macedonian language and the exclusions for Serbian, hoping to reach an accuracy of over 99%. The simplicity of the model allows for better generalization and faster fine-tuning, while still capturing the essential patterns of syllabification in both languages.

References

- [1] Walker, S. (2014). *Typography & language in everyday life: Prescriptions and practices*. Routledge.
- [2] Häikiö, T., Bertram, R., & Hyönä, J. (2016). The hyphen as a syllabification cue in reading bisyllabic and multisyllabic words among Finnish 1st and 2nd graders. *Reading and Writing*, 29, 159-182.
- [3] McCaskill, M. K. (1990). *Grammar, punctuation, and capitalization: a handbook for technical writers and editors* (Vol. 7084). National Aeronautics and Space Administration, Office of Management, Scientific and Technical Information Division.
- [4] Schlücker, B. (2023). Compounding and linking elements in Germanic. In *Oxford Research Encyclopedia of Linguistics*.
- [5] Mitreska, M., & Zdravkova, K. (2023, May). Syllable and Morpheme Segmentation of Macedonian Language. In *2023 46th MIPRO ICT and Electronics Convention (MIPRO)* (pp. 1113-1118). IEEE.
- [6] Zdravkova, K., & Kuzmanova, J. (2024). Sonority Based Syllabification of Macedonian and Serbian. *Technical editors*, 11.
- [7] Krstev, C. (1991). Serbo-Croatian hyphenation: a TEX point of view. *TUGboat*, 12(2), 215-223.
- [8] Vitas, D., Pavlović-Lažetić, G., & Krstev, C. (2007). About word length counting in Serbian (pp. 301-317). Springer Netherlands.
- [9] Radojičić, M., Lazić, B., Kaplar, S., Stanković, R., Obradović, I., Mačutek, J., & Leššová, L. (2019). Frequency and length of syllables in Serbian. *Glottometrics*, 45, 114-123.
- [10] A. Kovač, M. Marković M: A Mixed-principle Rule-based Approach to the Automatic Syllabification of Serbian. *Contributions to Contemporary History / Prispjevki za Novejšo Zgodovino*. 2019.
- [11] G. Clements: The sonority cycle and syllable organization, *Phonologica*, 63-76, 1988.
- [12] Domingos, P. (1999). The role of Occam's razor in knowledge discovery. *Data mining and knowledge discovery*, 3, 409-425.

Слоговна подела речи македонског и српског језика применом звучности

Катерина Здравкова, Јана Кузманова

Сажетак

Овај рад разматра процес слоговне поделе речи македонског и српског језика применом оригиналног приступа који се заснова на звучности фонема. Његова главна претпоставка је да звучност фонема расте од приступа према језгру, да би опет пала према одступу. Да би се конструисао аутоматски систем за ефикасну и тачну слоговну поделу речи, македонске и српске фонеме су категорисане у неколико типова, а звучност ових фонема је дефинисана да би се одредиле границе слогова.

Оба језика користе самогласнике и слоготворни сонант р као носиоце слога. Скуп језгара у српском језику проширен је и сонантима л и н који могу бити слоготворни. Самогласницима и сонантима је додељена највиша звучност. Сугласници су подељени у две категорије: звучни и беззвучни. Српски језик препознаје три додатне категорије: праскави, струјни и сливени, при чему звучност опада од праскавих према сливеним сугласницима.

Да би се одредила језгра и границе слогова, увели смо методу троструке разлике која израчунава разлику у звучности између суседних фонема. Она је позитивна само када је фонема језгро слога. Да би ово правило важило и у случају узастопних самогласника, између њих се додаје фиктивни консонант FC. Звучност FC и границе речи је 0.

Модел слоговне цртице су развијени у програмском језику Python, са прилагођавањима за македонски и српски језик на основу звучности.

Основни модел за македонски језик је постигао тачност од 88,70%. Погрешне поделе су се првенствено односиле на специјалне суфиксе ски, ство и ствен, који према правопису увек припадају истом слогу. Након прилагођавања, тачност подела у македонском језику се попела на 97,59%. Проблеми су укључивали руковање сложеним суфиксима и групама сугласника.

У српском језику, највећи изазов је било неслоготворно р, чија је звучност била смањена на 6. Овим побољшањем, коначна прецизност је достигла 98,68%.

Паралелно са сопственим приступом, упоредили смо резултате користећи ChatGPT. Прво смо га пажљиво обучили, а после првих

неуспешних покушаја, поновили смо обуку објашњавајући грешке. Упркос овим покушајима, ChatGPT је показао лошу тачност поделе речи на слоге (61,71% за македонски и 79,70% за српски), посебно због убацивања транслитерираних слова у неким речима, а каткад је правио и правописне грешке.

Приступ заснован на звучности који предлажемо показао се као једноставан, али ефикасан метод за поделу речи на слоге, који даје високу прецизност за оба језика. Очекујемо да би даља побољшања могла укључити информације у вези врсте речи за македонски језик и фино подешавање звучности за српски језик. Крајњи циљ ових модификација је да тачност подела за оба језика надмаши 99%.

Кључне речи: македонски, српски, слоговна подела, звучност фонема, ChatGPT.

Ученички корпуси србофоних говорника енглеског језика у контрастивној анализи међујезика на примеру језичке неодређености

Научни рад

DOI: 10.18485/judig.2025.1.ch6

Јелена Марковић¹  0000-0003-4962-3754

Апстракт

У овом чланку говоримо о контрастивној анализи међујезика из угла србофоне академске заједнице, односно описујемо постојеће ученичке корпусе србофоних говорника енглеског језика: *ICLE-SE*, компоненту тренутно најзначајнијег ученичког корпуса енглеског језика, *The International Corpus of Learner English v3* (UCLouvain, Белгија), као и *Корпус студената англистике (КорСАНг)* при Центру за дигиталну хуманистику Филозофског факултета Пале, користећи их на примеру истраживања феномена језичке неодређености.

Феномен језичке неодређености посматрамо доминантно кроз неодређено *some*, уз неодређене заменичке облике *somebody*, *someone*, *something*, односно еквивалентне форме неодређеног *неки* у српском језику. Резултати, које добијамо након анализе и тумачења коришћења циљних експонената језичке неодређености у писању на енглеском као страном у корпусу *ICLE-SE*, указују на учестало и снажно ослабање србофоних говорника на поменуте експоненте, неретко на уштрб јасноће или природности исказа. Даљим поређењима уочавамо да србофони говорници енглеског значајно чешће посежу за наведеним облицима од изворних говорника других матерњих језика. Анализирање употреба еквивалентних облика у писању на српском као матерњем језику на примерима корпуса *КорССАНг* уверава нас да

¹ Универзитет у Источном Сарајеву, Филозофски факултет Пале, jelena.markovic@ff.ues.rs.ba

постоји таква тенденција и у овом корпусу, како по фреквентности тако и по снажном ослањању на еквивалентно *неки*.

Закључујемо да је тенденција коришћења неодређеног језика у енглеском као страном код србофоних говорника резултат како универзалних тенденција у писању на страном језику тако и културолошких и језичких особености писања на српском као матерњем језику. Додатно закључујемо да је за сродна методолошка истраживања контрастивна анализа међујезика незаобилазна методологија у истраживањима ученичких корпуса, а поменути србофони ученички корпуси ресурси значајног потенцијала за сродне теме у примењеној лингвистици на српском говорном подручју.

Кључне речи: контрастивна анализа међујезика, *International Corpus of Learner English v3*, *КорСАНг*, језичка неодређеност

1. Увод

Појављивање контрастивне анализе међујезика (енгл. *Contrastive Interlanguage Analysis*),² као новог типа контрастирања у примењеној лингвистици, резултат је редефинисања и иновирања контрастивне методологије у теорији учења и усвајања страног језика. У новонасталом дигиталном окружењу овај нови тип контрастирања пратило је формирање првих обимних електронских ученичких корпуса (енгл. *learner corpora*), као нове парадигме података у истраживањима у оквиру примењене лингвистике (Granger, 2002; Gilquin, Granger, & Paquot, 2007), за разлику од до тада коришћене оскудне парадигме језичке продукције.³ Појава ученичких корпуса стога представља много више од основе за иновирање контрастирања, јер заправо представља нову фазу у развоју теорије учења и усвајања страних језика, која је за кратко време свог постојања дала велики научни допринос широм света: објављивање бројних монографских и других публикација, осмишљавање нових универзитетских курсева на различитим нивоима, извођење докторских истраживања, покретање нових научних часописа, уз значајан пратећи теоријски допринос у

² Термин *Contrastive Interlanguage Analysis* први пут је употребила Силвијана Грејндер [Sylviane Granger] 1996. године у раду под називом “From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora”.

³ Изненађујуће је колико је језичка продукција неизворних говорника до појаве ученичких корпуса заправо каскала у периоду процвата примењене лингвистике, очигледно занемарена у великом броју како методолошких тако и теоријских неусаглашености и изазова с којима се примењена лингвистика сусретала у својим првим деценијама формалног постојања (Chaudron, 2003: 774–776; Gass & Polio, 2014: 168–169; Марковић, 2019).

комплементарним дисциплинама попут филозофије језика, прагматике, анализе дискурса, социолингвистике и стилистике.

Критеријуми формирања електронских корпуса у циљу сврсисходности морају да буду строги и експлицитни. Наиме, морају се бележити примери језичке продукције најпре заједно са пратећим и релевантним метаподацима, а одлука о томе који ће метаподаци бити укључени зависи од проницљивости и расуђивања у предвиђању које се варијабле могу сматрати значајним за будуће истраживаче, што јесте захтеван корак у формирању електронских база аутентичне језичке продукције. Подразумева се да се метаподаци, без обзира на то које се варијабле бирају, бележе систематично и доследно, будући да једино такво бележење одабраних метаподатака има научни потенцијал. Метаподаци могу бити класификовани у више типова метаподатака, попут варијабли самих говорника (енгл. *learner variables*) или варијабли контекста језичке продукције (енгл. *task variables*). Као примере варијабли самих говорника наводимо узраст, односно старосну доб, дужину учења страног језика или познавање других страних језика. Као примере варијабли контекста језичке продукције наводимо циљни жанр, фактор могућности коришћења извора или речника, фактор интегрисаности говорне ситуације у испитне задатке или формално образовање или фактор временског ограничења говорне ситуације.

Нагласићемо да се критеријум одређивања дискурса као аутентичног у језичкој продукцији неизворних говорника неумитно редефинише у односу на одређивање аутентичности у језичкој продукцији изворних говорника. Наиме, за разлику од животног окружења у којем се свакодневно аутентичност говора изворног језика заправо подразумева, простор за апсолутну аутентичност у учењу страних језика у учионици веома је ограничен (Sinclair, 1996; Granger, 2002: 8). Стога се аутентичном продукцијом на страном језику може сматрати језичка продукција која има за циљ комуникацију на страном језику, а може се одвијати у условима који су делимично контролисани, како истиче и Елис (Ellis, 1994: 671). Другим речима референтност корпуса у истраживањима ученичких корпуса по критеријуму аутентичности постиже се у прихватљивом степену уколико се продукција може сматрати комуникацијом у одређеном контексту. С друге стране, пошто се намеће потреба контекстуализоване писане дискурсне продукције у ученичким корпусима, један од начина постизања циља у формирању корпуса јесте, на пример, писање састава, а не појединачних реченица.

Ревидирани методолошки модел контрастивне анализе међујезика (Granger, 2015: 17) иновиран је у односу на претходно дефинисани првенствено истицањем варијабилности, односно дефинисањем варијанте језика која се истражује, како у случају изворног тако и ученичког корпуса, с тим што се под варијантом језика може подразумевати свако специфично раслојавање у употреби језика.

2. International Corpus of Learner English v3

Ученички корпус *International Corpus of Learner English v3* (ICLE v3), формиран је при Католичком универзитету у Новом Лувену (фр. Université catholique de Louvain, Louvain-la-Neuve, скр. UCLouvain), односно Центру за англистичку корпусну лингвистику (*Centre for English Corpus Linguistics*) као трећа верзија корпуса, публикована 2020. године (Granger et al., 2020).⁴ Ова верзија садржи око 5,7 милиона речи, односно 25 подкорпуса одређених матерњим језиком говорника, међу којима је и српски језик.

На платформи која подржава корпус,⁵ интегрисани су одговарајући корпусни алати, тако да је коришћење платформе додатно олакшано. Корпус је иначе и морфосинтаксички тагиран, а приручник садржи детаљна упутства о коришћеном сету ознака за основну и напредну претрагу. Додаћемо да напредна претрага није само именом напредна, већ омогућава задавање веома различитих упита, континуираних и дисконтинуираних, што, уз све претходно наведено, пружа бројне истраживачке могућности. При том истичемо да упити нису увек и искључиво засновани на формалним карактеристикама језичких јединица, већ су често у питању комбинације формалних и семантичких, односно функционалних карактеристика упита, уз висок степен тачности резултата.

3. ICLE-SE, ученички корпус србофоних студената англистике

Један од подкорпуса који је интегрисан у поменути корпус ICLE јесте корпус србофоних говорника енглеског језика као страног, ICLE-SE, а формиран је захваљујући учешћу четири универзитета, односно четири катедре за енглески језик на српском говорном подручју (Источно Сарајево, Бања Лука, Београд, Нови Сад). Корпус се састоји од 325 састава просечне дужине од 623 речи, доминантно аргументативног типа, које су писали углавном студенти завршних

⁴ Прва верзија публикована је 2002. (Granger et al., 2002), а друга верзија 2008. године (Granger et al., 2008).

⁵ <https://corpora.uclouvain.be/cecl/icle/home>

година првог циклуса студија. Сваки учесник припремио је по један састав, што је омогућило формирање веома репрезентативне базе података.⁶

Корпус *ICLE-SE* формиран је углавном без временског ограничења, такође углавном без коришћења речника, приручника или других референтних издања, а у условима који нису подразумевали испитне задатке.

Сликовит опис тематског оквира корпуса *ICLE-SE* даје набрајање неколико најчешће коришћених наслова састава, редом од најпопуларнијег, а то су били следећи:

- Is reading a book better than watching its movie adaptation?
- One should never judge a person by external appearances.
- Most university degrees are theoretical and do not prepare students for the real world.
- Are we too dependent on computers?
- In the words of the old song "Money is the root of all evil." Agree or disagree.
- Conflicts are necessary for healthy relationships.
- Does age matter in relationships?
- Discuss the advantages and disadvantages of banning smoking in restaurants.
- In our modern world, dominated by science, technology and industrialization, there is no longer a place for dreaming and imagination. What is your opinion?
- Discuss the necessity of make-up.

4. Корпус КорСАНг

Корпус студената англистике (*КорСАНг*)⁷ формиран је при Филозофском факултету Универзитета у Источном Сарајеву, у периоду од 2016. до 2019. године (Радоња и Шућур, 2021). Настао је захваљујући сарадњи четири катедре за англистику на српском говорном подручју, у Источном Сарајеву, Бањој Луци, Београду и Новом Саду, у оквиру два национална пројекта при ресорном

⁶ Индикативно је истаћи и резултат који је остварен у оквиру оцењивања насумично одабраних 20 састава из сваког националног подкорпуса, а у складу с Заједничким европским оквиром (Common European Framework of Reference for Languages). Наиме, за корпус србофоних говорника, *ICLE-SE*, односно за његових 20 састава, 4 састава оцењена су као ниво B2, чак 11 састава оцењено је као ниво C1, а преосталих 5 састава као ниво C2, што је свеукупно резултат који вреди истаћи (Granger et al., 2020: 12).

⁷ <https://noske.rgf.rs/>

Министарству Владе Републике Српске.⁸ Корпус има око 140.000 речи, односно 327 текстова. У формирању корпуса укупно је учествовало 194 студената англистике, изворних говорника српског језика.

По саставу овај корпус је трокомпонентан. Садржи корпус превода са енглеског језика (*КорПСАнгл1*), корпус превода на енглески језик (*КорПСАнгл2*) и корпус аргументативних састава студената англистике писаних на српском језику као матерњем (*КорССАнгл*). Прве две компоненте јесу паралелни ученички корпуси, док је трећа компонента, *КорССАнгл*⁹ осмишљена као референтни корпус корпусу *ICLE-SE*, претходном описаном подкорпусу међународног ученичког корпуса аргументативних састава *ICLE* у трећој верзији.

Корпус *КорПСАнгл1* састоји се од 127 превода са енглеског на српски језик (око 47.200 токена). Полазни текстови на енглеском језику укључују и књижевне и некњижевне текстове. Један учесник припремио је највише два превода (једног књижевног и једног некњижевног текста). Сваки превод има свој код који означава правац превођења, универзитет, врсту текста, годину студија учесника и редни број превода за дати текст на том универзитету. Укупно је коришћено шест различитих текстова, који су преузети из више извора, углавном часописних за некњижевне текстове.

Корпус *КорПСАнгл2* састоји се од 130 превода са српског на енглески језик (око 51.000 токена). Полазни текстови на српском језику имају и књижевни и некњижевни карактер. Један учесник припремио је највише два превода: један књижевни и један некњижевни текст. Сваки превод има свој код, формиран по принципу као и за *КорПСАнгл1*. Укупно је коришћено шест различитих текстова из више извора, углавном часописних за некњижевне текстове.

Корпус *КорССАнгл* састоји се од 70 аргументативних састава просечне дужине око 570 речи (око 45.000 токена). Састави су писани на опште теме, тако да чине референтни корпус ученичком корпусу србофоних говорника *ICLE-SE*. Наслови састава следе, заједно с бројем састава који су писани на појединачне теме (Прилог 2):

⁸ Називи пројеката су „Фразеолошка компетенција српских говорника енглеског кроз призму контрастивне анализе међујезика“ и „Научни потенцијал анотираних ученичких корпуса кроз призму контрастивне анализе међујезика“, оба остварена уз подршку ресорног Министарства Владе Републике Српске (редом, број 19/6-020/961-46/18, од 31. 12. 2018. године и број 19.032/961-135/19, од 31. 12. 2019.).

⁹ Комплетан корпус *КорСАнгл* називамо ученичким у овом чланку, узимајући у обзир његов састав, као и могуће употребе у истраживањима. Прецизније би било описати компоненте превода ученичким корпусима, а компоненту *КорССАнгл* корпусом састава изворних говорника.

- Породица или посао: шта је важније? (9)
- Како музика утиче на живот? (8)
- Како замишљам доброг родитеља (7)
- Од склапања брака до развода данас (6)
- Сиромаштво и здравље (5)
- Да ли друштвене мреже нарушавају односе међу људима? (4)
- Ја (не) бих у Европску Унију (3)
- Пластична хирургија: да или не? (3)
- Млади не треба да се баве политиком (3)
- Да ли је брак превазиђена институција? (3)
- Зашто се гријеши при избору професије? (2)
- Зашто (не) желим да будем наставник страног језика (2)
- Слобода медија данас (2)
- Будућност књиге (2)
- Пушење на јавним мјестима: да или не? (2)
- Како замишљам идеално друштво (2)
- Љепота је пролазна (2)
- Родитељи би требало да поштују личност своје дјеце (1)
- Шта решавају затворске казне? (1)
- Значај образовања у друштву у коме живимо (1)
- Да ли критичко мишљење има своје мјесто у нашем образовном систему? (1)
- Без муке нема науке (1)

Сваки учесник припремио је по један аргументативни састав, без изузетка. На тај начин обезбеђена је максимална репрезентативност овог корпуса, иначе наизглед релативно малог обима у поређењу с другим корпусима. Сваки састав обележен је својим кодом.

Комплетан корпус *КорСАНг* аотиран је вишеструко. Најпре, урађена је библиографска аотација, која је повезала постојеће метаподатке по варијаблама са текстовима учесника. Даље, аотирани су сви текстови врстама речи (*PoS tagging*), а затим је урађено и поравнање паралелних корпуса превода на нивоу реченице (*sentence alignment*). Приликом обраде текста сачувано је писмо које су учесници користили у писању или куцању. У текстовима су отклоњене једино поједине очигледне техничке грешке у куцању, у циљу постизања ефективнијег претраживања корпуса. Изворни текстови представљени су у оригиналном облику, како су задати (без форматирања). Корисницима је омогућено претраживање корпуса путем корисничког интерфејса. Планирано је и проширење овог

корпуса укључивањем нових сродних текстова, по истим принципима формирања, посебно када је у питању трећа компонента, односно корпус референтан корпусу *ICLE-SE*.¹⁰

5. Феномен језичке неодређености

У оквиру овог чланка одабрали смо феномен језичке неодређености као студију случаја којом дајемо увид у могућности коришћења србофоних ученичких корпуса, *ICLE-SE* и *КорССАнг*, у примењенолингвистичким истраживањима.

Феномен језичке неодређености или неодређеног језика (енгл. *vague language*) представља једну од инхерентних особености природних језика, иако наизглед природни језици теже прецизности и одређености. Већ почетком двадесетог века Расел је забележио да је језик, уопштено говорећи, суштински мање или више нејасан или неодређен, управо због различите природе механизма природних језика и свеукупности значења која се изражавају (Russell, 1923). Нешто касније, Карл Попер [Carl Popper] је нагласио да се језик заправо суштински опире или отима потпуној прецизности, насупрот нашим очекивањима.¹¹

Феномену језичке неодређености у новијем периоду поклања се више пажње, првенствено у оквиру интердисциплинарних области анализе дискурса, примењене лингвистике и прагматике. Теорија језика, постало је извесно, уколико претендује да буде потпуна теорија, неизоставно укључује феномен језичке неодређености (Channell, 1994: 5). У најширем смислу, можемо рећи да је неодређени језик заправо резултат одступања од прецизности и јасноће у комуникацији (Jucker et al., 2003: 1737), али при том одмах треба истаћи да је коришћење неодређеног језика заправо типично у свакодневној употреби језика. Другим речима, језичка неодређеност само наизглед личи на непожељан феномен. У суштини, није у питању некакав недостатак или мањкавост, већ заправо незаобилазна компонента комуникативне, односно стратегијске и прагматичке компетенције којом се свакодневно служе изворни говорници (Carter & McCarthy 2006: 202; Parvaresh & Ahmadian, 2016: 437). Језичка

¹⁰ Користимо прилику да обавестимо научну јавност на српском говорном подручју да у предстојећем периоду планирамо проширивање корпуса састава на српском језику чији су аутори студенти, изворни говорници српског језика, примарно филолошке оријентације.

¹¹ Видоусон (Widdowson, 1984: 58) преноси мишљење Карла Попера, који каже да је трагање за прецизношћу ("the quest for precision") аналогно трагању за сигурношћу ("the quest for certainty"), а да је у језику немогуће испунити такве тежње, јер је идеја о прецизности у језику погрешна. Језик се опире прецизности због своје суштинске недоречености.

неодређеност заправо представља „напор који користимо попут стратегије у постизању циљева комуникације“¹² (Parvaresh & Tayebi, 2014: 597).

Неодређеност може бити изражена језичком формом која нема прецизно значење у датом контексту (Zhang, 2011). То може бити реч, синтагма или комплекснија језичка јединица. Неки од експонената у савременом енглеском језику као матерњем у свакодневној комуникацији јесу *thing, stuff, or so, like, something, anything, and so on, whatever, kind of, sort of* (Carter and McCarthy, 2006: 202).

Неодређеност језичких експонената реализује се у специфичном контексту, будући да значење истог експонента неодређености може бити другачије у другачијем контексту (Zhang, 2011: 572):

- There are *many* people in his room.
- There are *many* people in the park.

Некада је употреба неодређених језичких израза језика заправо једини избор говорника, јер се саопштава информација која је сама неодређена, а некада је то заправо намерни и вољни избор говорника који бира неодређени језик као релевантнији у датој говорној ситуацији, као што се истиче у савременим студијама (Jucker et al., 2003: 1737, McGee, 2018: 45):

- *Many* friends attended her birthday party. (Zhang, 2022: 274)

Many може имати различите интерпретације у зависности од више фактора: нпр. особе која слави (у смислу популарности слављенице, односно ширини круга пријатеља), површине простора за прославу или броја гостију које простор може да прими (Zhang, 2022: 274), што уједно показује да неодређени облик може бити прецизнији од навођења конкретног броја присутних у датој ситуацији.

Можемо резимирати наведено констатацијом да је коришћење неодређеног језика заправо манифестација прагматичке компетенције говорника. Употреба језичких експонената неодређеног значења у језику мотивисана је и сврсисходна, тако да представља индикатор вештине (изворног) говорника, односно промишљено језичко понашање у интеракцији (Jucker et al., 2003: 1739).

¹² “an endeavor strategically made to achieve certain communicative goals” (превод аутора)

6. Неодређеност у продукцији на страном језику

Једна од основних карактеристика међујезика јесте његова варијабилност, видљива у језичкој продукцији неизворних говорника (Saville-Troike, 2006). Паралелно с израженим степеном варијабилности, неизворни говорници у покушају да надоместе своје језичке недостатке, користе различите стратегије, односно стратегијску компетенцију, значајно више од изворних говорника, који имају на располагању језичку компетенцију и познавање релевантних социолингвистичких, културолошких и прагматичких феномена.

Међу препознатљивим стратегијама неизворних говорника налази се и когнитивна тенденција ка генерализацији, односно коришћењу уопштенијих, неодређених језичких форми како у превођењу на страни језик (Spajić & Suknović, 2019) тако и у писању на страном језику (Hasselgren, 1994; Lee & Chen, 2009). И само разумевање неодређеног језика битно је различито код изворних и неизворних говорника. Стога можемо истаћи да је веома захтевно за неизворне говорнике да овладају коришћењем различитих дискурских карактеристика страног језика, укључујући и прагматички оправдано коришћење језичке неодређености (Evison et al., 2007; Parvaresh & Ahmadian, 2016; Marković, 2019; Marković, 2021), посебно имајући у виду комплексност значења и употреба израза неодређености у изворном језику.

7. Неодређени језик у корпусу ICLE-SE

У наставку следе најпре примери употребе неодређеног језика из србофоног корпуса *ICLE-SE*, остварени коришћењем различитих језичких механизма:

- However, there are *some* cases in which this *kind of* an obstacle is certainly relevant. SEBJ1016¹³
- *Sometimes*, we all get tired, or there are *some* days when we are not pretty to ourselves, so why not put *some* make-up on. (SEBJ1004)
- Just because *someone* looks a *certain* way it doesn't mean they are good or bad. (SENS1021)
- There are *some* movies that all of us love and we can watch them over and over again, [...]. (SEBJ1011)
- It can happen to *anyone* to fall in love with *someone* who is older or younger, or *someone* who is different or *whatever*. (SEBJ1022)

¹³ Након појединачних примера, назначен је код састава из корпуса.

Понављање језичких јединица неодређеног значења, различитих или некада чак истих, веома је често у оквиру исте реченице код србофоних говорника, чиме се стиче утисак неприродности, односно неизворности дискурса:¹⁴

- Long, long time ago, *someone* very conservative and narrow-minded said that, in order to do a *certain* job you have to look a *certain* way. (SENS2001)
- Unlike *many*, *some* companies realised *how many* problems smoking can cause for a *large number of* people, so they put warning messages on every cigarette pack like <*> or <*> and so on. (SEBJ1003)

У фокусу овог чланка јесу првенствено неодређено *some*, уз неодређене заменичке облике *somebody*, *someone*, *something*, који су и одабрани као експоненти језичке неодређености у корпусу србофоних говорника *ICLE-SE*, а које називамо циљним неодређеним облицима у наставку текста.

У корпусу србофоних говорника учесталост наведених облика¹⁵ релативно је висока (0,87%), што је и очекивано и разумљиво будући да је у питању писање на енглеском као страном језику. Ипак, поређење с ученичким корпусима говорника других страних језика из истог међународног корпуса *ICLE* (укупно 18, видети Табелу 1., Прилог) показује да је код говорника српског језика, уз турски, учесталост већа од говорника других матерњих језика (бугарског, чешког, кинеског, холандског, финског, француског, немачког, грчког, мађарског, италијанског, литванског, македонског, персијског, пољског, руског, шпанског и шведског), као и од процентуалне заступљености на нивоу комплетног корпуса (0,53%) и референтног дела корпуса *LOCNESS* (0,39%), иначе корпуса изворних говорника (Granger, 1998).¹⁶ Другим речима, иако, са једне стране, постоји општа тенденција коришћења израза језичке неодређености код неизворних говорника, с друге стране уочавамо да је присуство циљних неодређених облика учесталије управо у корпусу србофоних говорника, што је и предмет нашег истраживања.

¹⁴ Приликом формирања коментара о наведеним примерима употребе из корпуса затражили смо и мишљење лектора, изворног говорника америчке варијанте енглеског језика.

¹⁵ Задата претрага укључила је *some* и све сложене неодређене облике чија је прва морфема *some*.*.

¹⁶ Корпус састава изворних говорника који се типично користи као референтни корпус у истраживањима због сличних социолингвистичких особености говорника формиран при Центру за англистичку корпусну лингвистику у Новом Лувену (Centre for English Corpus Linguistics, Université catholique de Louvain, Louvain-la-Neuve). (<https://www.learnercorpusassociation.org/resources/tools/locness-corpus/>).

Из конкорданци које су генерисане коришћењем интегрисаних корпусних алата на платформи корпуса *ICLE* издвојили смо неколико група примера употребе. Најпре наводимо примере који оправдано и прагматички јасно користе неодређени облик *some*, чак и поновљен у оквиру исте реченице:

- *Some* adaptations are successful, and *some* are not, mainly because *some* movies are successful, and *some* are not. (SEBG2021)
- *Some* people are against big age differences, *some* are not. (SEES2012)
- *Some* film adaptations are hits, and *some* are misses, but those hits hit hard. (SEBG2021)
- *Some* people that are under twenty seem to be more mature than *some* people that are over that age. (SEBJ1031)
- *Some* people tend to hide it, *some* are not ashamed to admit that it can't all be rainbows and butterflies, and *some* just don't give it too much thought. (SEBG2004)
- Пример употребе других неодређених облика следи:
- The question is how long we will wait until *someone* will do *something* serious to improve education system in Serbia. (SEBG2001)

Поједини примери нису неразумљиви, али је чињеница да коришћење неодређених облика одузима флуентност и захтева пажњу приликом читања:

- For instance, if *someone* sees a man with considerably younger woman the first thing that comes to their mind is that he is probably an extremely wealthy businessman and that finances are *the thing* that brought them together. (SEBG1026)¹⁷
- In the end we should not reject movie adaptations completely, because *sometimes* we can learn *something* new from them. (SEBJ1014)

За разлику од претходних, чести су примери где је неразумљивост евидентна, с тим што је управо коришћење неодређеног језика извор нејасноће. Такве примере било би потребно преформулисати у циљу постизања јасноће:

¹⁷ У корпусу *ICLE* задржана је потпуна аутентичност текстова неизворних говорника, укључујући и неправилности у формама речи (нпр. *financies).

- Even if *someone* is not trying to be *someone* they are not by their appearance, how they like to dress does not have to represent their personality. (SEBG2017)¹⁸
- Secondly, *everyone*, based on personal preferences and characteristics, is looking for *someone* who suits them, for *someone* it is an older person, and for *someone* younger. (SEBJ1016)¹⁹
- Generally, *some* people are more likely to judge *other* people immediately after they meet *someone* for the first time than *other* people. (SEBJ3005)²⁰

У следећим примерима непотребно оптерећење представља управо неодређено *some*. Стога, оправдано га је изоставити из примера који следе:

- But quite often movies that are adaptations of *some* books are not as good as the books. (SEBJ1027)
- In some cases, as *some* people are with their clothes, so they are with their habits. (SENS3014)
- Money is largely responsible for the enormous economic growth of *some* countries like America, Russia, China, Australia etc., where the utopian dream proposed by many a scholar has been realized. (SEBG2048)
- Native speakers are able to acquire some new words, or to add *some* new meanings to already known words. (SEES2026)
- What I remember best is that the invention of money immediately triggered *some* conflicts among people. (SENS1027)
- When *some* people hear where you're coming from, they make a picture of you and your country [...].(SEBJ1039)
- Yes, I still haven't lost hope in the movie industry and I hope there will be *some* better adaptations of this everlasting story in the future. (SEBG1028)

У примерима који следе уместо неодређеног *some* боље је користити неодређени члан, *a*:

¹⁸ Even if someone is not trying to express themselves through their appearance, how they like to dress does not have to represent their personality. (SEBG20172)

Напомена: Предлоге свих наведених парафраза у фуснотама, које су обележене индексом 2, припремио је изворни говорник енглеског језика америчке варијанте.

Конкретно за наведени пример изворни говорник је назначио да је извор нејасноће и језичка неодређеност и двострука негација у реченици.

¹⁹ Secondly, everyone, based on personal preferences and characteristics, is looking for someone who suits them; for one, it might be an older person, and for another, someone younger. (SEBJ10162)

²⁰ Generally, some people are more likely to judge other people immediately after they meet someone for the first time. SEBJ3005²

- Secondly, descriptions given in books never tell you something certain; for example, it is up to you how you will imagine *some* character's physical appearance, his/her facial expressions or the way he/she moves. (SEBG2030)
- Somebody says there is *some* boundary of five years difference and that should be ok. (SEBJ1031)
- All things considered, we can say that reading *some* book is always better than watching a movie. (SEES1008)

У примерима који следе уместо неодређеног *some* боље је користити неодређени придев *certain* или *various* (+pl.) у завршној реченици:

- Firstly, there are some universities that do not have the appropriate subjects for *some* degrees, so students are obliged to study subjects that are not so closely related to their major. (SEES2008)
- Age is not a crucial thing when you have to discuss *some* issues. (SEBJ1031)
- There are *some* social conventions that some people tend to follow. (SEBJ1050)
- [W]e learn to value different culture by being taught on different aspects from *some* country, such as literature, customs, its history, etc. we develop our vocabulary, listening, writing and speaking skills. (SEBG2051)²¹

8. Неодређени језик у корпусу КорССАнг

Значај формирања корпуса *КорССАнг*, који садржи око 45.000 речи или 70 аргументативних састава писаних на српском језику чији су аутори изворни говорници српског, може се видети управо у поређењу које представљамо. Наиме, након запажања да су циљни неодређени облици учесталији код србофоних говорника, приступили смо анализи изворног корпуса *КорССАнг* и претражили неодређено *неки* у њему.²² Напомињемо да ниједан од наслова састава коришћених за формирање корпуса не садржи употребе циљних неодређених облика, што би у противном могло бити доведено у везу с учесталим употребама. Резултат обављене претраге јесу укупно 264

²¹ [W]e learn to value different cultures by being taught different aspects from various countries, such as literature, customs, its history, etc.; we develop our vocabulary, listening, writing and speaking skills. SEBG2051²

²² [lemma_lc="неки"]

примера употребе, односно учешће с 0,58% заступљености.²³ Одабрали смо неколико примера.

Најпре, има примера употребе, попут следећих, које можемо оценити као јасне и разумљиве изворним говорницима:

- Kada *neko* kaže „heroj“, prva asocijacija na tu riječ je moja majka. (KopCCAn232)
- Kao da je znanje *neki* kolač koga ima u određenoj količini, pa što ga više ljudi deli, to manje svako pojedinačno dobija. (KopCCAn263)
- Npr. meni se često dešava da čujem *neku* pjesmu koja ima neprikladan ili totalno besmislen tekst, ali veoma dobar ritam. (KopCCAn234)
- Npr. *neka* osoba sluša određenu pjesmu kad raskine sa svojim partnerom. (KopCCAn234)
- Krv, novac, estrada, i *neke* ekstremno bitne informacije poput, gdje spava Cecin sin večeras. (KopCCAn23)
- [I]mamo slučaj da *neke* zemlje (Njemačka, Francuska...) imaju potpisan sporazum o bezviznom ulasku u neke od najrazvijenijih zemalja svijeta (SAD, Kanada) [...]. (KopCCAn215)

У наредним примерима умножавање циљних неодређених облика може представљати когнитивно оптерећење и за изворног говорника:

- Поготово када питате *неког* човјека за кога је гласао на изборима, а он вам одговори да није изашао на изборе уопште. Такви људи нису *никоме* јасни. Како *неко* може да са таквом лакоћом пусти *некога* да му сутра одлучује о судбини. (KopCCAn259)
- [„] Banja Luka idealno mjesto za druženje, socijalizaciju“, i to jeste (donekle) tačno. Na svakom koraku možeš vidjeti *nekoga*. Sa *nekim/a*. (KopCCAn236)
- Ukoliko *neko* nije zadovoljan *nekim* dijelom svog tijela, ili želi *nekakvu* promjenu na sebi kako bi se osjećao zadovoljnije [...]. (KopCCAn21)
- Zašto *nekoga* osuđivati i zamjerati mu *nešto*? (KopCCAn212)
- [...] jer je sve u *nekoj* mjeri važno za *nekoga* i za *nešto*. (KopCCAn223)
- Kako mi reagujemo na *neku* pjesmu mnogo zavisi od stanja našeg uma ili *nekih* prethodnih iskustava. Ako *neko* određeno razdoblje svog života povežemo s *nekom* pjesmom [...]. (KopCCAn234)

У примеру који следи у недоумици смо да ли су несигурни и они који не воле свој изглед иста група људи, или је ипак реч о две различите групе:

²³ Упит CQL [lemma_lc="неки|нечији|неки|нечији" & tag="PRO"] даје 273 резултата, односно 0,6%.

- *Neki* ljudi su nesigurni u sebe, *neki* ne vole svoj izgled, nešto im smeta i žele da se promene. (KopCCAn230)
- Примере који следе радо бисмо парафразирали, опет због циљних неодређених облика:
- *Неки* од разлога зашто би требало бити забрањено је тај што дим цигарета може сметати *некоме*. (KopCCAn258)
- Из тог разлога, можемо тврдити да музика делује искључиво терапеутски, као утеха у звуку који [је] створен од стране *некога* ко је имао иста или слична искуства, *неко* ко разуме. (KopCCAn250)
- *Niko nikome* nije naudio što se zaljubio pa želi da se razvede. To je u redu. Međutim, nije u redu ako *taj neko* nije pošten pa ne prizna to i nastavi da živi u bračnoj zajednici [...].(KopCCAn212)

У примерима који следе можемо замислити исте реченице, али без неодређеног *неки*. На пример, у првој реченици јасно је да *неки* посао у просвети наставнички позив, односно да је довољно рећи „посао у просвјети“:

- Уколико је једини посао који могу да пронађем у својој струци *неки* посао у просвјети, онда би се искрено радије запослила у *неком* кафићу или butikу него радила као наставник. (KopCCAn218)
- Слично је и у бројним примерима који следе:
- [...] ljudi treba da posjeduju *neku* pomoć od ljudi koji imaju moć da jednim pozivom riješe problem. (KopCCAn25)
- Ako su već u tim fazama da žele *neku* promjenu, uvijek mogu promijeniti frizuru, boju kose, način odjevanja. (KopCCAn27)
- [...] jednostavno su presebični da se odreknu nečega zarad druge osobe, zarad *nekog* većeg, opšteg dobra. (KopCCAn28)
- Ulazak u brak uliva *neko* osećanje sigurnosti pa se ljudi polako opuštaju i pokazuju svoje pravo lice ili se jednostavno ulenje [...]. (KopCCAn29)
- U prošlosti su ovakvi problemi bili manje izraženi jer su se više njegovalе *neke* vrijednosti poput solidarnosti i poštovanja [...]. (KopCCAn219)
- Trebamo da se odlučimo, hoće li to biti porodica, посао или ће се пак покушати ostvariti *neki* balans između to dvoje, i oprobati se u obe te sfere života. (KopCCAn27)
- Ponekad muzički žanr koji preferiramo može da kaže mnogo o našoj ličnosti i to je razlog za *neke* smiješne ali pak česte predrasude koji ljudi imaju. (KopCCAn233)

Навођењем неколико група примера из корпуса *KopCCAn2* нисмо исцрпили све маштовите употребе неодређеног *неки*.

9. Дискусија и закључна разматрања

Чињеница је да је коришћење израза неодређености у писању на енглеском као страном у значајном степену универзална тенденција код неизворних говорника. Ипак, код србофоних говорника, као што смо видели, коришћење циљних форми, *some*, *somebody*, *someone*, има своје учестале, неоправдане или неприхватљиве реализације. Другим речима, постојећа навика ослањања на поменуте неодређене облике у писању на енглеском језику код србофоних говорника често има слабо прагматичко, односно комуникативно упориште.

Претрага и анализа еквивалентног неодређеног облика *неки* у корпусу *КорССАн* показала је изненађујуће сличне тенденције, како квантитативно тако и квалитативно. Наиме, осим високе учесталости у коришћењу циљних форми у писању и на страном и на матерњем језику, србофони говорници снажно се ослањају на неодређене облике и у матерњем језику у ситуацијама када то није неопходно, некада више пута у оквиру једне реченице, а некада им радије прибегавају у поређењу с алтернативним формама сличног функционалног домена.

На основу наведеног, намеће се закључак да је коришћење неодређеног експонента *неки* код србофоних студената заправо најпре тенденција у писању на матерњем језику, самим тим и рефлексција друштвених и културолошких фактора. Поменута тенденција преноси се на продукцију на енглеском као страном, додатно оснажена универзалним карактеристикама често неоправданог или претерано присутног степена неодређености.

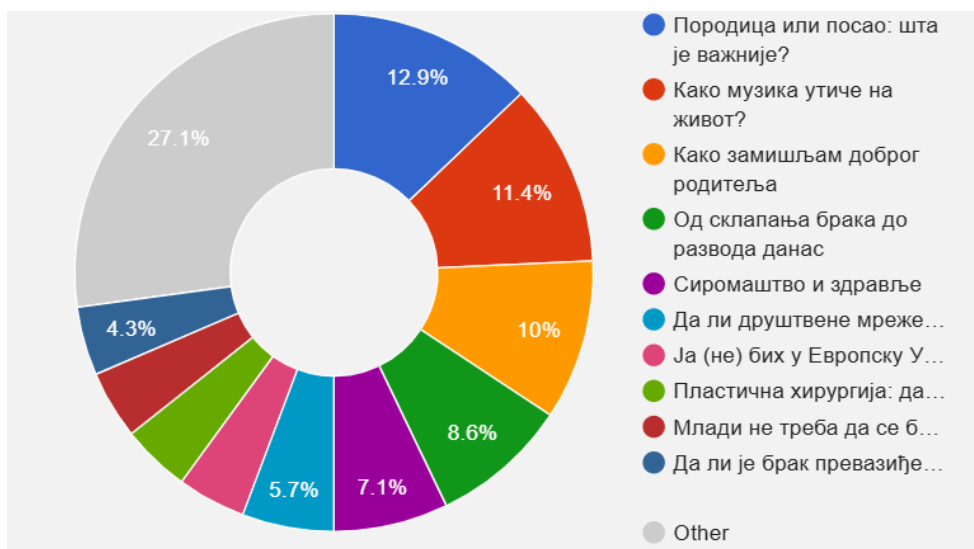
Извесно је да би истраживање требало даље проширити варијантама српског као матерњег код говорника различитих социолингвистичких варијабли (првенствено старосне доби и образовања), будући да се закључци до којих смо дошли не могу генерализовати на говорнике српског језика уопште. За такво проширивање истраживања неопходно би било користити и доступне корпусе српског језика као матерњег, уз формирање подкорпуса за одговарајуће варијанте језика.

Значај формирања два корпуса које смо користили у истраживању постаје јаснији на примеру изведене студије. Надамо се и даљем проширивању корпуса *КорССАн*, будући да би свако проширење било драгоцено за откривање и објашњавање различитих дискурних феномена у међујезику србофоних говорника, што даље носи са собом научне и педагошке импликације.

Табела 1. Учесталост неодређених облика у различитим корпусима

	<i>some</i> * (%)	<i>some</i> (%)
<i>ICLE</i>	0.53	0.32
<i>LOCNESS</i>	0.39	0.23
Serbian	0.87	0.42
Turkish	0.87	0.43
Persian and Urdu	0.66	0.46
Czeck	0.62	0.32
Chinese	0.60	0.52
Bulgarian	0.58	0.32
Finnish	0.58	0.31
Russian	0.54	0.30
Macedonian	0.53	0.33
Swedish	0.48	0.24
Polish	0.47	0.31
Lithuanian	0.46	0.29
Spanish	0.43	0.24
German	0.42	0.21
Greek	0.42	0.20
Italian	0.42	0.21
Dutch	0.41	0.23
French	0.41	0.26
Hungarian	0.30	0.19

ПРИЛОГ 2



Графикон 1. Приказ учесталости најпопуларнијих наслова састава у корпусу *KorCSAнг*

Извори

- [1] <https://noske.rgf.rs/>.
- [2] <https://www.learnercorpusassociation.org/resources/tools/locness-corpus/>.
- [3] <https://corpora.uclouvain.be/cecl/icle/home>.

Литература

- [1] Carter, R., & McCarthy, M. (2006). *Cambridge Grammar of English*. Cambridge: Cambridge University Press.
- [2] Channell, J. (1994). *Vague Language*. Oxford: Oxford University Press.
- [3] Chaudron, C. (2003). Data collection in SLA Research. In C. J. Doughty, & M. H. Long (Eds.), *Handbook of Second Language Acquisition* (pp. 762–828). Oxford: Blackwell Publishing.
- [4] Ellis, R. (1994). *The Study of Second Language Acquisition*. Oxford: Oxford University Press.
- [5] Evison, J., McCarthy, M., & O’Keeffe, A. (2007). 'Looking out for love and all the rest of it': Vague category markers as shared social space. In J. Cutting (Ed.), *Vague Language Explored* (pp. 138–157). Basingstoke: Palgrave MacMillan.
- [6] Gilquin, G., Granger, S., & Paquot, M. (2007). Learner corpora: The missing link in EAP pedagogy. *Journal of English for Academic Purposes*, 6(4), 319–335.
- [7] Gass, S., & Polio, C. (2014). Methodological influences of "Interlanguage" (1972): Data then and data now. In Y. Han, & E. Tarone (Eds.), *Interlanguage: Forty years later* (pp. 147–172). Amsterdam and Philadelphia: John Benjamins.
- [8] Granger, S. (1996). From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. In K. Aijmer, B. Altenberg, & M. Johansson (Eds.), *Languages in Contrast: Text-based cross-linguistic studies* (pp. 37–51). Lund: Lund University Press.
- [9] Granger, S. (1998). The computer learner corpus: A versatile new source of data for SLA research. In S. Granger (Ed.), *Learner English on Computer*. London and New York: Addison Wesley Longman.
- [10] Granger, S. (2002). A Bird's-eye view of learner corpus research. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching* (pp. 3–33). Amsterdam and Philadelphia: John Benjamins.
- [11] Granger, S. (2015). Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research*, 1(1), 7–24.

- [12] Granger, S., Dagneaux, E., & Meunier, F. (Eds.). (2002). *The International Corpus of Learner English. Handbook and CD-ROM*. Louvain-la-Neuve: Presses universitaires de Louvain.
- [13] Granger, S., Dagneaux, E., Meunier, F., & Paquot, M. (Eds.). (2008). *The International Corpus of Learner English. Handbook and CD-ROM (Version 2)*. Louvain-la-Neuve: Presses universitaires de Louvain.
- [14] Granger, S., Dupont, M., Meunier, F., Naets, H., & Paquot, M. (Eds.). (2020). *The International Corpus of Learner English v3*. Louvain-la-Neuve: Presses universitaires de Louvain.
<https://dial.uclouvain.be/pr/boreal/object/boreal:229877>
- [15] Hasselgren, A. (1994). Lexical teddy bears and advanced learners: A study into the ways Norwegian students cope with English vocabulary. *International Journal of Applied Linguistics*, 4, 237–258.
- [16] Jucker, A. H., Smith, S. W., & Ludge, T. (2003). Interactional strategy: Interactive aspects of vagueness in conversation. *Journal of Pragmatics*, 35, 1737–1739.
- [17] Lee, D. Y., & Chen, S. X. (2009). Making a bigger deal of the smaller words: Function words and other key items in research writing by Chinese learners. *Journal of Second Language Writing*, 18, 281–296.
- [18] Marković, J. (2019). “It is a thing that gives you...”: The lexeme thing(s) as the Serbian EFL ‘teddy bear’. *Komunikacija i kultura online*, 10, 19–37.
- [19] Marković, J. (2021). I, you, and we in Serbian EFL Argumentative Writing from the Essay Title Perspective. *Folia Linguistica et Litteraria*, 36, 271–290. doi:10.31902/fl.36.2021.16
- [20] Parvaresh, V., & Ahmadian, M. (2016). The impact of task structure on the use of vague expressions by EFL learners. *The Language Learning Journal*, 44, 436–450.
- [21] Parvaresh, V., & Tayebi, T. (2014). Vaguely speaking in Persian. *Discourse Processes*, 51, 565–600.
- [22] Russell, B. (1923). Vagueness. *Australian Journal of Philosophy and Psychology*, 1, 84–92.
- [23] Saville-Troike, M. (2006). *Introducing Second Language Acquisition*. Cambridge: Cambridge University Press.
- [24] Sinclair, J. (1996). EAGLES. Preliminary recommendations on Corpus Typology. Retrieved November 5, 2018, from www.ilc.pi.it/EAGLES/corpus/corpus.html

- [25] Spajić, S., & Suknović, M. (2019). The Choice of Lexemes According to their Frequency in Translation into L2. *Komunikacija i Kultura Online*, 10, 104–119. doi:10.18485/kkonline.2019.10.10.6
- [26] Widdowson, H. G. (1984). *Explorations in Applied Linguistics 2*. Oxford: Oxford University Press.
- [27] Zhang, G. (2011). Elasticity of vague language. *Intercultural Pragmatics*, 8(4), 571–599. doi:10.1515/IPRG.2011.026
- [28] Zhang, G. Q. (2022). Vague Language from a Pragmatic Perspective. In I. Kecskes (Ed.), *In The Cambridge Handbook of Intercultural Pragmatics* (pp. 274–300). Cambridge: Cambridge University Press.
- ***
- [29] Марковић, Ј. (2019). Кроз призму контрастивне анализе међујезика. Пале: Филозофски факултет.
- [30] Радоња, М. и Шућур, С. (2021). О Корпусу студената англистике (КорСАНГ) и могућностима његове софтверске експлоатације. *Infothesa – Journal for Digital Humanities*, 21(37–58).

Serbian EFL Learner Corpora in Contrastive Interlanguage Analysis Through The Example of Vague Language

Jelena Marković

Summary

This paper focuses on the two Serbian EFL learner corpora, i.e., ICLE-SE and KorSANG, exploring their potential for applied linguistics research in the field of Second Language Acquisition through the example of Vague Language exponents. The learner corpus ICLE-SE is a Serbian subcorpus of The International Corpus of Learner English, version 3 (ICLE, Université catholique de Louvain, Belgium), currently the most significant learner corpus of English. The KorSANG corpus consists of three parts: the two parallel corpora of English-Serbian and Serbian-English translations, and a subcorpus of argumentative essays in L1 Serbian, collected with the purpose to serve as a referential corpus for ICLE-SE.

The focus of the article is the use of Vague Language, as an inherent language feature exploited on a daily basis. In actual fact, Vague Language has proved to be a pragmatically justified means of language use, fulfilling various contextual purposes, only seemingly resembling a language

imperfectness or deficiency. The two learner corpora have been used in order to compare and contrast the use of the targeted indefinite forms, some→*, and neki.

The research is embedded within Contrastive Interlanguage Analysis, a renowned contrastive methodology in Second Language Acquisition. The research has shown that Serbian EFL writing exploits the targeted Vague Language more frequently than the referential native corpus; moreover, Serbian EFL writing contains the targeted forms more frequently than EFL writing originating from several other L1 backgrounds. The analysis of the KorSAng subcorpus serving as the ICLE-SE referential, has shown that native speakers of Serbian use the Vague Language form neki almost as frequently in their L1 Serbian as they use its English equivalents in their EFL writing. Moreover, there are striking similarities in both L1 and EFL writing, i.e., overreliance through repetition, optional or sometimes erroneous uses, and unclear utterances.

We conclude that the use of the targeted Vague Language forms in EFL writing results from both universal features of EFL writing and the L1 writing habits of EFL writers. Vague Language turns out to be challenging not only for Discourse Analysis, but even more for Applied Linguistics, since it is an extremely demanding feature of the foreign language system. We also conclude that the two learner corpora in question possess a huge potential in Second Language Acquisition research for cognate research questions regarding Serbian EFL written production.

Key words: Contrastive Interlanguage Analysis, International Corpus of Learner English v3, KorSAng, Vague Language

Deep Learning versus Rule-based Models: Is Rule-based Methodology a Thing of the Past? A Study of Hotel Review Sentiment Analysis

Scientific paper

DOI: 10.18485/judig.2025.1.ch7

Maram Alharbi^{1,2},  0009-0004-7028-9214

Ruslan Mitkov^{1,3}  0000-0003-2522-066X

Abstract

In the rapidly evolving domain of sentiment analysis within the hospitality sector, accurately understanding customer sentiment from hotel reviews is increasingly crucial for market intelligence and improving service quality. This study focuses on categorising hotel reviews into positive and negative sentiments using a range of sentiment analysis models. We examine both rule-based approaches—such as VADER, AFINN, and TextBlob—and deep learning methods, including the T5 transformer model and BiLSTM. While rule-based models employ predefined sentiment lexicons, offering simplicity and interpretability, they often fall short in capturing the complexity of emotional nuances and context-dependent sentiments. Conversely, deep learning models, particularly those leveraging advancements in transfer learning, provide a more sophisticated understanding of linguistic intricacies, enabling more effective sentiment detection, even in complex reviews. Our experiments show that rule-based models achieve moderate performance, with F1 scores of 0.79 for TextBlob, 0.77 for AFINN, and 0.78 for VADER. While efficient, these methods often struggle with the contextual subtleties of sentiment. In contrast, deep learning models demonstrated superior performance, with the T5 transformer model achieving an F1 score of 0.96 and the BiLSTM model reaching 0.93. This highlights the potential of deep

¹ Lancaster University

² Jazan University

³ University of Alicante

learning techniques for sentiment analysis in customer reviews, offering more accurate and nuanced sentiment classification than traditional methods. We compare these approaches across two datasets of hotel reviews, assessing the impact of various preprocessing techniques and sentiment analysis models in terms of accuracy, precision, and recall. Our findings emphasise the superior performance of deep learning models, particularly the T5 transformer, in accurately classifying sentiments and addressing the challenges posed by mixed sentiments within reviews. Nonetheless, rule-based models retain their utility in scenarios where computational efficiency is prioritised. This study provides a comprehensive evaluation of sentiment analysis methods within the hospitality industry, offering insights that can enhance both customer experience and business strategy. By comparing traditional and advanced sentiment analysis techniques, we contribute to a deeper understanding of model performance and their practical applicability in real-world settings.

1. Introduction

Sentiment Analysis (SA), also known as opinion mining or emotion AI, is a key application of Natural Language Processing (NLP) focused on understanding and categorising emotional undertones within digital communications. Its primary objective is to discern and evaluate subjective insights conveyed through text [1]. Emerging in the early 2000s alongside the proliferation of online content [2], sentiment analysis has become essential in interpreting emotional tones in digital communications.

In the hospitality industry, online consumer reviews are invaluable for enhancing business performance. User-generated reviews on digital platforms significantly influence customer decision-making, with many potential guests consulting reviews before booking hotels. Insights from these reviews reflect collective opinions and directly impact travellers' choices and decisions [3].

The advent of advanced NLP models has significantly benefited sentiment analysis, providing nuanced insights into customer feedback. This field ranges from basic linguistic processing to contextual interpretation, enabling hotel management to better understand customer sentiment and improve service quality [4].

This study seeks to compare the performance of rule-based methods and deep learning models for sentiment analysis. While deep learning models are expected to perform better, we argue that it is important to formally validate this premise all the more so because other recent studies [5] have shown that rule-based approaches are not to be underestimated and

deliver better results than a range of deep learning models. More specifically, we experiment with the rule-based sentiment analysis systems VADER, AFINN and TextBlob, comparing their capabilities to address our research goals. These lexicon-based models use predefined sentiment lexicons for a straightforward approach to sentiment analysis. However, their effectiveness can be limited by domain-specific nuances [6].

While rule-based approaches are generally effective, they often fall short in capturing the complex, context-dependent nature of sentiment. Deep learning approaches have emerged as a potent alternative, capable of learning intricate patterns and contextual nuances essential for accurate sentiment analysis. However, these models are not infallible and can underperform in certain areas, such as translating multiword expressions or anaphora resolution [7, 8, 9, 10, 11].

Advancing our analytical framework, this study incorporates deep learning models T5 transformer and BiLSTM for sentiment analysis in the hospitality industry. The use of transfer learning and transformer models streamlines the application of pre-trained models to specific tasks. This study also assesses the deep learning models' capabilities in interpreting complex sentiment dynamics within textual narratives.

The rest of the paper is structured as follows. Section 2 surveys related work in the field. Section 3 presents the methodology adopted including the data used for this study, the data processing stages, and outlines the rule-based approaches and deep learning model employed. Section 4 reports the evaluation results, while section 5 discusses the results and offers error analysis. Finally, section 6 summarises the conclusion of this study.

2. Related work

Rule-based models in sentiment analysis use predetermined rules and lexicons to decipher the emotional tone of text. These models rely on a predefined set of words, each associated with specific sentiment values, to evaluate and categorise textual data. Their simplicity and straightforward methodology make them effective in various applications, including customer feedback analysis in the hospitality sector [3]. VADER (Valence Aware Dictionary and sEntiment Reasoner) and AFINN are lexiconbased methods used for sentiment analysis, effective at identifying complex sentiment patterns but limited in grasping context or subtle sentiment expressions [6], while TextBlob's performance can be influenced by the nature of the textual data and preprocessing steps [12].

The evolution of sentiment analysis has been influenced by deep learning models. While neural networks were proposed in the 1990s, the impact of Deep Learning on NLP, became particularly pronounced after 2015 [13]. This period marked a transformative era in the adoption of neural network methodologies for complex linguistic tasks.

Raffel et al. [14] introduced an innovative approach in NLP, treating various language tasks uniformly as text-to-text problems. Their T5 model transformed all text-based problems into a text-to-text format, facilitating the application of the same model across diverse tasks such as summarisation, question answering, and text classification with minimal task-specific adaptations.

Following this, Pipalia et al. [15] delved into sentiment analysis within NLP, evaluating the effectiveness of various transformer-based models. Their research particularly emphasised the proficiency of the BiLSTM in understanding long-term dependencies in text, an essential aspect for nuanced sentiment analysis.

Lastly, Pathak et al. [16] analysed the transformative impact of transformer-based models, particularly accentuating the advancements in learning capabilities brought about by these models. The paper highlighted T5's unique approach of treating all NLP tasks as text-to-text conversions, integrating features from previous models and displaying very high performance results, including near-human scores in some benchmarks.

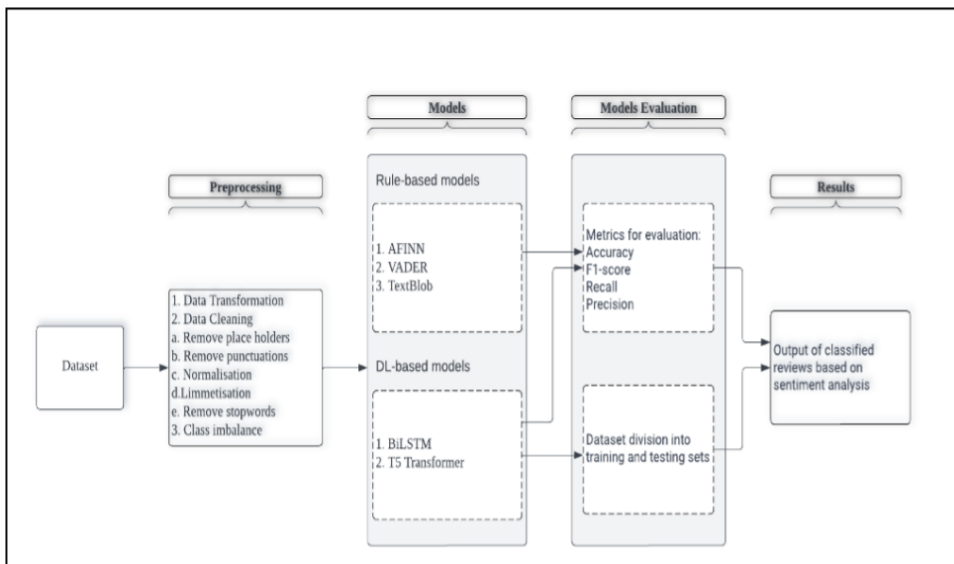


Figure 1: Sentiment Analysis Methodology Framework

3. Methodology

Figure 1 showcases the framework of our study, illustrating the structured approach to sentiment analysis applied to hotel reviews.

3.1. Data Selection and Description

This study utilises two publicly available datasets from Kaggle.

- **515K Hotel Reviews Dataset:** Contains 515K customer reviews from 1,493 luxury hotels across Europe. Attributes include hotel address, review date, nationality, detailed feedback, average score, reviewer tags, and geographical coordinates. Reviews are categorised as 'Positive' or 'Negative' reviews.
- **38K Hotel Reviews Dataset:** Comprises 38,932 reviews for a single hotel. Attributes include detailed reviews, browser and device used, and categorisation as 'positive' or 'negative' reviews.

3.2. Data Preparation

Key steps in data preparation include:

- **Removing Placeholder Text:** Excluding phrases like 'No Positive' and 'No Negative' that do not provide meaningful insights.
- **Normalisation:** Converting text to lowercase, removing URLs, punctuation, handling repeated letters, and removing stop words.
- **Lemmatisation:** Identifying each word's base form to enhance semantic accuracy.

This study primarily focuses on reviews written in English.

3.3. Rule-Based Methods

VADER, AFINN, and TextBlob are rule-based sentiment analysers which categorise hotel reviews into positive or negative sentiments. VADER converts raw text into sentiment scores, which are then classified as positive or negative. AFINN uses a lexicon-based approach, assigning sentiment scores to words on a scale from -5 to +5 to determine overall sentiment. TextBlob provides a straightforward method for sentiment extraction, making it effective at evaluating sentiments in hotel reviews.

3.4. Deep Learning-based Models

3.4.1. Bidirectional Long Short-Term Memory (BiLSTM) Model

The BiLSTM model is used for sentiment analysis, classifying hotel reviews as positive or negative. It processes input data in both forward and backward directions, capturing the sequence and context of words.

The initial step involves tokenising the text into words or phrases and converting these tokens into numerical indices. This allows the model to perform numerical operations on the text data. The model uses embeddings to transform these indices into high-dimensional vectors, capturing the lexical and semantic relationships between tokens. For example, "hotel" and "motel" are represented by vectors placed near each other in the vector space due to their similar meanings. To ensure uniform input sequence lengths, shorter sequences are padded with zeros, and longer ones are truncated. The BiLSTM model processes these sequences bidirectionally, enhancing its ability to understand complex sentence structures and nuanced language use, making it suitable for sentiment analysis tasks.

Training the BiLSTM model involves tuning its architecture and parameters. The architecture includes an Embedding layer for converting integer sequences into dense vectors, a BiLSTM layer for capturing long-term dependencies and contextual nuances, and a Dense output layer with a sigmoid activation function for binary classification of sentiments. The model's hyperparameters are optimised for performance: 1000 maximum words for the tokenizer, 50-dimensional embedding vectors, 128 BiLSTM units, and a binary cross-entropy loss function with the Adam optimiser. Training is conducted over 20 epochs with a batch size of 32, using early stopping to prevent overfitting.

3.4.2. Text-To-Text Transfer Transformer (T5) Model

The T5 model, a Transformer-based neural network, was used in this study to categorise hotel reviews into sentiment classes through tokenisation, model training, and sentiment classification.

The T5 Tokenizer processes the text, converting it into tokens and mapping each to a unique input ID. An attention mask is also created, directing the model's focus to relevant tokens while ignoring the rest. The T5 model is fine-tuned for sentiment analysis with a learning rate of $5e-5$, batch size of 32, and 10 training epochs. The Adam optimiser and CrossEntropyLoss function are used to minimise prediction errors. The maximum sequence length is set at 512 tokens to maintain consistency. The

T5-Small variant balances complexity and performance, fitting computational constraints while delivering robust sentiment analysis.

Finally, the models' performance is evaluated using two datasets, ensuring robust validation. The evaluation metrics used are precision, recall, F1-scores, and accuracy. The BiLSTM and T5 models, designed to capture contextual nuances, are tested on new reviews to verify their applicability. Their predictions are decoded into sentiment labels ('positive' or 'negative').

4. Evaluation Results

The comparative analysis of the sentiment analysis models' performance employed traditional metrics and confusion matrices for a deeper understanding of model behaviour in sentiment classification tasks. Additionally, we evaluated the impact of stop word removal on model performance by comparing accuracy and F1 scores before and after the exclusion of stop words.

To ensure the reliability of our findings, the evaluation of the deep learning models was conducted over multiple runs. Each model's performance metrics were averaged across these runs, with standard deviations reported to account for variability due to random initialisations and training conditions.

Table 1: Comparative Performance Metrics of Sentiment Analysis Models Across Two Datasets in Experimental Evaluations

Dataset	Models	Accuracy	F1 Score		Recall		Precision	
			Pos	Neg	Pos	Neg	Pos	Neg
Dataset A	VADER	0.71	0.78	0.48	0.99	0.33	0.64	0.95
	AFINN	0.69	0.74	0.49	0.99	0.33	0.59	0.96
	TextBlob	0.76	0.79	0.71	0.84	0.66	0.75	0.76
	BiLSTM	0.93	0.93	0.93	0.92	0.94	0.94	0.92
	T5	0.96	0.96	0.96	0.95	0.97	0.97	0.95
Dataset B	VADER	0.69	0.76	0.55	0.96	0.39	0.63	0.91
	AFINN	0.71	0.77	0.62	0.92	0.49	0.66	0.85
	TextBlob	0.62	0.73	0.35	0.99	0.21	0.57	0.96
	BiLSTM	0.85	0.85	0.85	0.82	0.88	0.88	0.82
	T5	0.89	0.90	0.89	0.91	0.87	0.89	0.90

This approach provides a more robust representation of the models' performance across the two datasets. The performance of the rule-based and deep learning models was assessed using accuracy, F1 scores, recall, and precision. These metrics offered a nuanced view of each model's predictive capabilities.

Building on the analysis presented in Table 1, a comprehensive comparison of various models is provided using metrics such as accuracy, precision, recall, and F1 score. These metrics enable a critical evaluation of each model's strengths and weaknesses, offering a robust basis for comparing rule-based and deep learning approaches. The results highlight that deep learning models like T5 and BiLSTM deliver more consistent performance across datasets, with T5 achieving an accuracy and F1 score of 0.96 on the larger dataset, followed by BiLSTM with an accuracy of 0.93.

In contrast, the performance of rule-based methods varied depending on dataset size and class balance. TextBlob demonstrated better results on the larger dataset, while VADER and AFINN excelled in precision for negative sentiment classification. On the smaller dataset, T5 continued to lead, though with a slight decrease in accuracy, and BiLSTM's performance dipped, particularly in recall for negative sentiments. AFINN showed improvements in positive sentiment classification, while VADER and TextBlob faced declines in precision and recall.

The focus on class-specific metrics for positive and negative sentiments ensures a more direct and clear comparison between models. This approach allows for a precise evaluation of each model's ability to classify sentiment accurately, aligning the results closely with the practical requirements of sentiment analysis tasks. The analysis is grounded in quantitative evidence, reinforcing the validity of the study's conclusions.

Figure 2 presents an analytical comparison of F1 scores for the models, evaluating the impact of stop word inclusion versus exclusion. Incorporation of stop words typically resulted in enhanced F1 scores for the detection of positive sentiments across most models. Conversely, the removal of stop words was associated with a uniform reduction in F1 scores for negative sentiment detection.

Figure 3 describes the accuracy of the selected models, quantifying their performance in terms of the proportion of sentiments correctly classified out of the total observations. The T5 and BiLSTM models demonstrated consistency in accuracy, with minor variations observed following the removal of stop words.

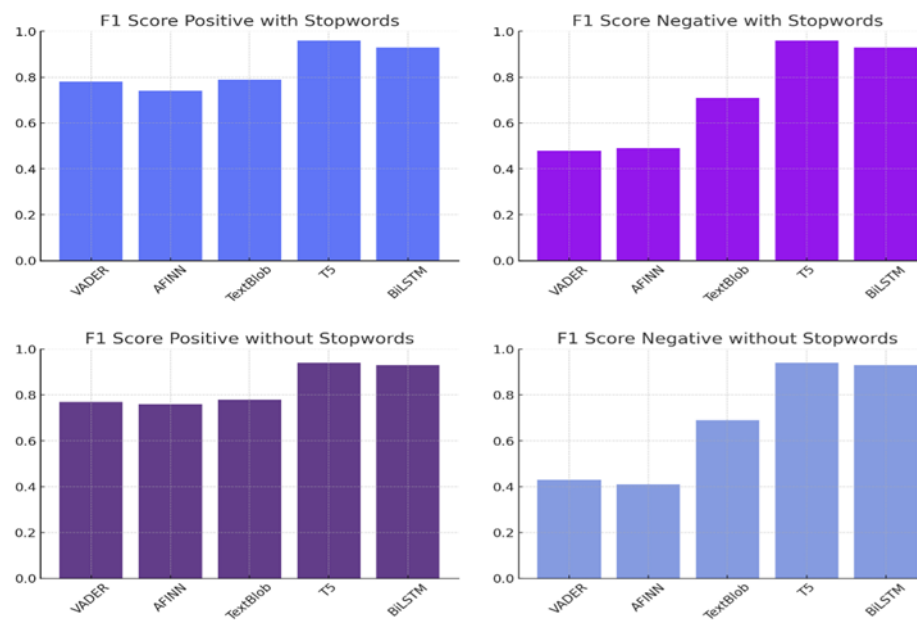


Figure 2: Models F1 Score With vs. Without Stop Word

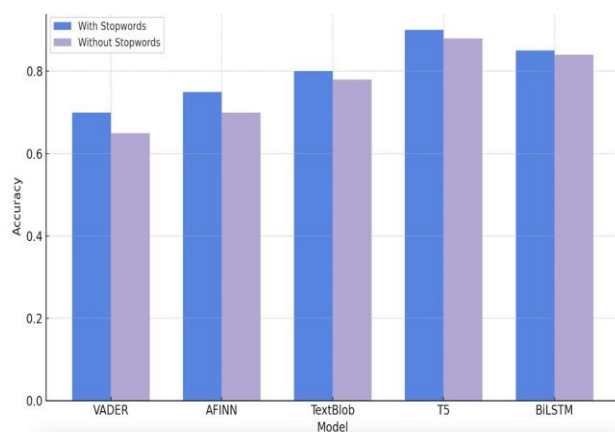


Figure 3: Models Accuracy With VS. Without Stop words

5. Discussion

The varying performance of sentiment analysis models underscores the significant impact of dataset characteristics on model efficacy. Dataset A, with shorter and more straightforward reviews, simplifies sentiment classification. In contrast, the longer and more complex reviews in Dataset B introduce challenges but also opportunities to demonstrate analytical capabilities.

The rule-based models struggled with the intricacies of the longer reviews in Dataset B, showing a decline in performance. This suggests difficulties in adapting to linguistic diversity and subtle sentiments. However, AFINN showed some improvement on Dataset B, likely due to the frequent use of words in its sentiment lexicon. The removal of stop words during preprocessing further affected the performance of rule-based models, as these words carry essential contextual information. Advanced models like T5 and BiLSTM were less impacted by the removal of stop words, as they capture broader context and semantic relationships.

The T5 model excelled in contextualising and interpreting varied sentiments within a single narrative, while the BiLSTM model effectively processed sequential data, capturing sentiment nuances. Both models demonstrated the transformative impact of deep learning in sentiment analysis, with T5 setting a new benchmark in the field. Despite the superior performance of deep learning models in context-based decision-making, rule-based models remain relevant due to their efficiency, simplicity, and lower computational requirements.

Notably, our analysis revealed that rule-based models, particularly VADER and AFINN, can outperform deep learning models in scenarios with less contextual complexity or when computational resources are limited.

5.1. Error and Comparative Analysis of Sentiment Analysis Models

Performance errors across sentiment analysis models often arise due to the complex nature of review texts, leading to incorrect classifications. Each case of errors is quantified by calculating precision, recall, and F1 scores for misclassifications specific to that error category. These measures are extracted by evaluating the performance of each classifier on a test set specifically curated to include representative examples of the error cases. These errors can be categorised into several specific cases, allowing for a comparative analysis of how different models handle these challenges.

Case 1: Handling of Misspellings

Misspellings present a challenge for both deep learning and rule-based models, affecting sentiment detection accuracy. For example: - "room was quit small and not very clena" can obscure the negative sentiment about the room size. - "service was excelent very freindly staff" may result in positive sentiment being overlooked.

Rule-based models, struggle with misspellings due to their reliance on direct lexical matches. VADER and TextBlob show moderate resilience but can miss unconventional spelling errors. Deep learning models handle misspellings better due to their advanced contextual analysis capabilities.

Case 2: Complexity of Mixed Sentiments in Reviews

Mixed sentiments within a single review require models to balance contrasting feedback, such as: - "the staff was friendly but the room was dirty" needs accurate weighing of positive and negative sentiments. - "great location, but the bed was uncomfortable" can lead to prioritising one sentiment over another.

Rule-based models often falter in accurately discerning overarching sentiment amid contradictory expressions. In contrast, the T5 model excels in evaluating and balancing contrasting sentiments within a narrative, showcasing its nuanced comprehension and classification of complex emotions.

Case 3: Incorrect Sentiment Labels

Training data with incorrect sentiment labels can mislead models, affecting classification accuracy. For example: - "disappointed with the noisy room" labelled as positive can skew model interpretations. - "hallway need redone" labelled as positive may misinterpret dissatisfaction as commendation.

Deep learning models effectively correct mislabelled sentiments, showcasing their adaptability and nuanced interpretation of sentiment.

Despite the superior performance of deep learning models, rule-based models occasionally outperform them in terms of computational efficiency and speed. Our analysis revealed that rule-based models excel in processing shorter reviews and handling explicit sentiment expressions. This suggests that hybrid methods, combining rule-based and deep learning approaches, could be advantageous in contexts where both computational efficiency and nuanced sentiment interpretation are required, such as in real-time sentiment tracking applications.

6. Conclusion

Our study focused on the hospitality sector, analysing hotel reviews to evaluate the effectiveness of traditional rule-based models versus advanced deep learning models in sentiment classification.

Our findings highlighted the superiority of the T5 transformer model in interpreting complex sentiments. However, rule-based models such as VADER, AFINN, and

TextBlob remain valuable in scenarios requiring limited computational resources or swift deployment.

The comparative analysis enriches the discourse in sentiment analysis and offers actionable insights for hotel management and the broader NLP community. Our evaluation confirms the effectiveness of deep learning models in capturing complex sentiment expressions, demonstrating their superiority over rule-based models.

7. Future Work

Future research will focus on aspect-based sentiment analysis to identify specific service or facility attributes that influence guest sentiments. Large Language Models (LLMs) and advanced deep learning technologies which could enhance sentiment analysis accuracy, will be also experimented with and compared. Additionally, crosslingual models are essential to reflect the hospitality industry's global nature and diverse guest experiences. Expanding language coverage and employing sophisticated algorithms will provide a more comprehensive understanding of sentiment analysis, ultimately improving guest experiences and driving industry success.

References

- [1] R. Mitkov, The Oxford Handbook of Computational Linguistics 2nd edition, Oxford University Press, 2022.
- [2] E. Demir, M. Bilgin, Sentiment analysis from turkish news texts with bert-based language models and machine learning algorithms, in: 2023 8th International Conference on Computer Science and Engineering (UBMK), IEEE, 2023, pp. 01–04.
- [3] A. Ameer, S. Hamdi, S. Ben Yahia, Sentiment analysis for hotel reviews: A systematic literature review, ACM Comput. Surv. 56 (2023). URL: <https://doi.org/10.1145/3605152>. doi:10.1145/3605152.
- [4] S. Mutmainah, D. H. Fudholi, Leveraging bilstm and lda for analyzing and dashboarding user feedback in applications, JURNAL MEDIA INFORMATIKA BUDIDARMA 8 (2024) 51–61.

- [5] R. Mitkov, L. A. Ha, Are rule-based approaches a thing of the past? the case of anaphora resolution, *Procesamiento del Lenguaje Natural* 72 (2024).
- [6] C. Hutto, E. Gilbert, Vader: A parsimonious rule-based model for sentiment analysis of social media text, in: *Proceedings of the international AAAI conference on web and social media*, volume 8, 2014, pp. 216–225.
- [7] J.-P. Colson, Multi-word units in machine translation: why the tip of the iceberg remains problematic—and a tentative corpus-driven solution, *Computational and Corpus-based Phraseology* (2019) 145.
- [8] Mitkov, Ruslan, Computer vs. human intelligence. keynote speech at the refinitiv conference, city of london., 2019.
- [9] R. Stuckardt, Machine-learning-based vs. manually designed approaches to anaphor resolution the best of two worlds. *proceedings of the discourse anaphora and anaphora resolution colloquium, daarc'4*, 211-216. lisbon, portugal., 2002.
- [10] R. Stuckardt, Three algorithms for competenceoriented anaphor resolution. *proceedings of the discourse anaphora and anaphora resolution colloquium, daarc'5*, 157-163. sao miguel, portugal., 2003.
- [11] R. Stuckardt, A machine learning approach to preference strategies for anaphor resolution. in antônio branco, tony mcenery, and ruslan mitkov (eds.), *anaphora processing: Linguistic, cognitive and computational modeling*, 47-72. john benjamins, amsterdam/philadelphia, 2004.
- [12] G. Sreenivas, K. M. Murthy, K. Prit Gopali, N. Eedula, M. H R, Sentiment analysis of hotel reviews - a comparative study, in: *2023 IEEE 8th International Conference for Convergence in Technology (I2CT)*, 2023, pp. 1–9. doi:10.1109/I2CT57861.2023.10126445.
- [13] . M. R. Orasan, C., Recent Developments in Natural Language Processing. In R. Mitkov (Ed.), *The Oxford Handbook of Computational Linguistics* 2nd edition. Oxford University Press., Oxford University Press, 2021.
- [14] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, *CoRR abs/1910.10683* (2019). URL: <http://arxiv.org/abs/1910.10683>. arXiv:1910.10683.
- [15] K. Pipalia, R. Bhadja, M. Shukla, Comparative analysis of different transformer based architectures used in sentiment analysis, in: *2020 9th International Conference System Modeling and Advancement in Research Trends (SMART)*, 2020, pp. 411–415. doi:10.1109/SMART50582.2020.9337081.
- [16] A. Pathak, et al., Comparative analysis of transformer based language models, in: *CS & IT Conference Proceedings*, volume 11, CS & IT Conference Proceedings, 2021.

Poređenje pristupa zasnovanog na pravilima i dubokog učenja za razumevanje osećanja

Maram Alharbi, Ruslan Mitkov

Sažetak

U domenu analize zadovoljstva ugostiteljskim uslugama koja se brzo razvija, tačno razumevanje osećanja i mišljenja korisnika iz recenzija hotela je sve važnije za inteligentno upravljanje tržištem i poboljšanje kvaliteta usluge. Ovo istraživanje se fokusira na kategorizaciju recenzija hotela na pozitivne i negativne komentare koristeći niz modela analize osećanja. Ispitujemo pristupe zasnovane na pravilima—kao što su VADER, AFINN i TektBlob—, kao i metode dubokog učenja, uključujući model T5 transformatora i BiLSTM. Dok modeli zasnovani na pravilima koriste unapred definisane leksikone osećanja, nudeći jednostavnost i interoperabilnost, oni često ne uspevaju da uhvate složenosti emocionalnih nijansi i osećanja zavisnih od konteksta. Nasuprot tome, modeli dubokog učenja, posebno oni koji koriste napredne transformerske modele, pružaju sofisticiranije razumevanje jezičkih kompleksnosti, omogućavajući efikasnije otkrivanje osećanja, čak i u složenim kontekstima. Naši eksperimenti pokazuju da modeli zasnovani na pravilima postižu umerene performanse, sa F1 rezultatima od 0,79 za TektBlob, 0,77 za AFINN i 0,78 za VADER. Iako efikasne, ove metode se često bore sa kontekstualnim suptilnostima osećanja. Nasuprot tome, modeli dubokog učenja pokazali su superiorne performanse, pri čemu je model transformatora T5 postigao F1 rezultat od 0,96, a BiLSTM model je dostigao 0,93. Ovo naglašava potencijal tehnika dubokog učenja za analizu sentimenta u recenzijama kupaca, nudeći precizniju i nijansiraniju klasifikaciju osećanja od tradicionalnih metoda. Upoređujemo ove pristupe u dva skupa podataka recenzija hotela, procenjujući uticaj različitih tehnika prethodne obrade i modela analize sentimenta u smislu tačnosti, preciznosti i odzivu. Naši nalazi naglašavaju superiorne performanse modela dubokog učenja, posebno transformera T5, u preciznoj klasifikaciji osećanja i rešavanju izazova koje predstavljaju mešana osećanja u komentarima. Bez obzira na to, modeli zasnovani na pravilima zadržavaju svoju korisnost u scenarijima gde je računarska efikasnost prednost. Ovo istraživanje pruža sveobuhvatnu procenu metoda analize sentimenta u hotelijerstvu, nudeći uvide koji mogu poboljšati i korisničko iskustvo i poslovnu strategiju. Upoređivanjem tradicionalnih i naprednih tehnika analize osećanja,

doprinosimo dubljem razumevanju performansi modela i njihove praktične promenljivosti u realnim okruženjima.

Ključne reči: Analiza osećanja, pristup zasnovan na pravilima, duboko učenje, transformeri, BiLSTM, T5, VADER, TextBlob, AFINN

Towards Classifying Predicates Involving Change

Scientific paper

DOI: 10.18485/judig.2025.1.ch8

Ivelina Stoyanova,  0000-0003-3771-435X

Svetlozara Leseva¹  0000-0001-8198-4555

Abstract

The paper discusses the semantic features of verbs of causal change which determine their realisation in text. The description framework is based on the frames from FrameNet and the shallow classification they provide in terms of verbs' semantics and the syntactic realisation of the corresponding configurations of frame elements. The classification covers four classes: 1) Verbs of change of physical integrity, shape, general condition and/or functionality; 2) Verbs of change of (measurable) inherent or acquired properties; 3) Verbs of creation; 4) Other verbs of change through processing or manipulation.

The main focus in the study is placed on the entity undergoing the change and the property which characterises the change. The analysis is supplemented with annotated examples from corpora.

Keywords: verb semantics, frame semantics, verbs of change, FrameNet, WordNet

1. Introduction

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

¹ Department of Computational Linguistics, Institute for Bulgarian Language, Bulgarian Academy of Sciences, iva@dcl.bas.bg, zarka@dcl.bas.bg

subsuming different types of dynamic predicates, which, under certain conditions may be interpreted as *activities* in the sense of Vendler (1967).

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

In the paper, we focus particularly on causative verbs of change in WordNet and the FrameNet semantic frames describing them, but the observations regarding the core frame elements are largely applicable to the inchoative counterparts of the respective frames. Moreover, we exclude large semantic classes such as verbs of motion, verbs of placing, verbs of emotional or psychological change, verbs involving animate objects, and some other classes which have their own specifics and warrant a separate study.

We analyse the property undergoing the change and the relevant semantic restrictions as grounds for the classification of the frames. Based on the outlined semantic features, a shallow classification within the semantic class of verbs of change is offered, aiming to cover the diversity within the class and to propose an approach to model their syntactic behaviour. As the classification is derived from the FrameNet frames, it groups together verbs with similar conceptual structure and syntactic properties. In particular, the shared invariant semantics of the predicates (in terms of the semantic frame they evoke) determines to a large degree the syntactic realisation of their participants. Our observations on the syntactic realisation of verbs of change are based on empirical material extracted from various corpora for English and Bulgarian.

The paper is organised as follows. After a brief discussion of the related works and the representation of verbs of change in lexical-semantic resources, we present a shallow classification of the verbs supplied with illustrative examples. The final section draws conclusions and raises some theoretical questions which outline directions for future work.

2. Related Work

Semantic classifications of predicates and the description of semantic relations between predicates and their arguments have been

undertaken within various theoretical approaches, based on: (a) verbs' syntactic properties and behaviour (Levin 1993; Pinker 1989; Goldberg 1994, etc.); (b) thematic structure (Chafe 1970; Longacre 1976; Van Valin and LaPolla 1997); (c) frame semantics (Fillmore 1982).

One of the most widely acknowledged and probably the largest-scale classification of English verbs has been proposed by Levin (1993). The author has divided verbal predicates into classes on the basis of their invariant semantics and the syntactic diatheses in which they participate. This work has served as a foundation for the description of English predicates in VerbNet (Kipper-Schuler 2005), a lexical-semantic resource presenting a shallow hierarchy of verb classes in the English language, the semantic roles describing the argument structure of the predicates from each class, the selective restrictions imposed on the arguments, their syntactic realisation, etc.

As exemplified by VerbNet, large-scale classifications of verbs (possibly along with other parts of speech) are devised as part of the semantic description of lexical units in semantic resources. WordNet (see next section) presents the lexis in a taxonomically organised semantic network whose nodes represent synonym sets. VerbAtlas (Di Fabio et al. 2019) enriches the description in the WordNet-inspired multilingual semantic resource BabelNet by assigning each synonym sets a frame corresponding to its prototypical predicate-argument structure described using semantic roles (along with relevant semantic restrictions defined over them).

FrameNet (Fillmore et al. 2003; Ruppenhofer et al. 2016, among others) is a lexical resource that couches the semantic and syntactic properties of lexical units in the apparatus of frame semantics (Fillmore 1982). It provides a robust conceptual description of lexical items in terms of the fragment of knowledge, or frame, they evoke when used in language.

A more detailed description of FrameNet and WordNet is presented in the next section, which elaborates on the verbs that are in the focus of this paper, namely verbs of change.

3. Verbs of Change in Language Resources

WordNet, also the Princeton WordNet or PWN (Miller 1995; Fellbaum 1998), is a large lexical database that represents comprehensively conceptual and lexical knowledge in the form of a network whose nodes denote cognitive synonyms (synsets) linked by means of a number of conceptual-semantic and lexical relations such as hypernymy, meronymy, antonymy, etc. Each synset is supplied with a gloss, possibly with usage

examples as well as with notes (if appropriate) on the grammatical, stylistic or other properties of the synset members. In addition to the Princeton WordNet, we use the Bulgarian WordNet, or BulNet (Koeva 2021), which is aligned with PWN at the synset level using unique synset identifiers.

WordNet provides a coarsely-grained semantic division in terms of a set of language-independent semantic primitives (semantic classes) assigned to all the nouns and verbs in the resource. The verbs fall into 15 groups, such as *verb.change* (verbs describing change in terms of size, temperature, intensity, etc.), *verb.cognition* (verbs of mental activities or processes), *verb.motion* (verbs of change in the spatial domain), *verb.communication* (verbs describing communication and information exchange), etc. Verbs of change in WordNet largely belong to the semantic class *verb.change*.

The causative verbs of change are predominantly organised in the WordNet subtree stemming from eng-30-00126264-v {*change, alter, modify*} ‘cause to change; make different; cause a transformation’ which covers 2,536 synsets. 46% of them are labelled as *verb.change*. However, some verbs of change fall into different semantic classes which specify their semantics (including the changing component) in a more concrete fashion, e.g. *verb.emotion* describes change in the emotional state, *verb.contact* covers change as a result of physical contact between objects, *verb.body* denotes change in the bodily experience or appearance, etc. A small group of relevant verbs are also found in the subtrees eng-30-01617192-v {*make, create*} ‘make or cause to be or to become’ and eng-30-01850315-v {*move, displace*} ‘cause to move or shift into a new position or place, both in a concrete and in an abstract sense’.

FrameNet (Baker et al. 1998; Baker 2008) is a lexical semantic resource that couches lexical and conceptual knowledge using the apparatus of frame semantics. Frames are conceptual structures that describe types of objects, situations, or events along with their components (frame elements) (Baker et al. 1998; Ruppenhofer et al. 2016). Depending on their status, the frame elements (FEs) are divided into core, peripheral, and extra-thematic (Ruppenhofer et al. 2016: 23–24). The core FEs, which would be the primary focus below, instantiate conceptually necessary components of a frame, which in their particular configuration make a frame unique and different from other frames.

FrameNet frames are organised into a hierarchical network, using a number of frame-to-frame relations (Ruppenhofer et al. 2016: 81–84). The following two relations (along with some others) determine the hierarchical internal structure of thematic verb classes: *Inheritance* – defined as a relationship between a parent frame and a more specific (child) frame, such

that the child frame elaborates on the parent frame, and *Using* (also called ‘weak inheritance’) – a type of relationship between two frames where the first one makes reference in a very general kind of way to the structure of a more abstract frame (used exclusively in instances where a part of the scene evoked by the child frame refers to the parent frame).

Verbs of change are covered by a shallow hierarchy of frames stemming from *Transitive action*, describing causative change, and *Transition to a state*, characterising non-causative change.

Our analysis relies on existing mappings between WordNet and FrameNet which assign FrameNet frames onto WordNet synsets (Leseva and Stoyanova 2020).

FrameNet also offers a set of annotated examples for lexical units evoking the corresponding frames (Ruppenhofer et al. 2016: 7–8). The annotation includes the verb, the frame elements and the syntactic components through which the frame elements are realised. The annotation provides information both about the explicit and the implicit frame elements (Ruppenhofer et al. 2016: 28–29). The empirically grounded linguistic generalisations about the syntactic realisations of frame elements are particularly valuable not only in the study of the target language (English) but as a point of departure for making observations cross-linguistically. We adopt the principles of annotation from the FrameNet annotated corpus and apply them in the annotation of examples in Bulgarian. A detailed conception of a FrameNet-modelled description of Bulgarian verbs is laid out in Koeva (2010, 2020).

4. Methodology

In order to collect the dataset for analysis, we rely on the existing mapping of FrameNet frames to WordNet synsets. The methodology is based on the following steps.

1) We select a set of FrameNet frames which describe verbs of change such that: a) evoke causative frames inheriting from the abstract frame *Transitive action*; b) the change occurring in the affected entity involves a particular attribute of the entity (Patient or other), thus excluding very generally-specified frames such as *Cause change*. Other exceptions are also made, as specified in the Introduction.

At this step we have identified 40 FrameNet frames inheriting directly or indirectly from *Transitive action*. The internal organisation of the specified sets of frames is illustrated in the classification in the following section.

2) We select synsets to which any of the analysed frames are assigned, resulting in a total of 239 synsets covering 569 verbs in English and 925 verbs in Bulgarian (the greater number is mostly due to the fact that aspect is a lexical property of the verbs and both the perfective and the imperfective members of the aspectual pairs are included as separate literals in the synset).

3) We compile a set of examples collected from corpora by performing automatic extraction and manual selection based on the relevance of the instances, followed by manual annotation. The annotation includes identification of frame elements and labelling them with the relevant syntactic category (NP.Ext, NP. Obj, PP, AdvP, etc.). The examples and annotations for English are borrowed from the FrameNet corpus.

The analysis focuses in particular on the affected entities and their properties with the further aim of defining classes with similar semantic characteristics and syntactic realisation in terms of the morphosemantic features of the verb and the configuration of frame elements (valency frames) it appears with.

In some cases where there is no frame in FrameNet that describes the studied verbs (usually groupings of synsets in a WordNet tree), we define such a frame with its corresponding frame elements and relations. Some of these frames, labelled with an asterisk, are integrated in the classifications presented in Figures 1 – 4.

5. Classification of Frames Representing Verbs of Change

Below we attempt to classify the causative verbs of change in terms of the characteristic features of the entities affected by the change and the property which is subjected to change or the state that occurs as a result of the change. The most general distinction in this respect is between 1) frames that involve affecting a Patient in such a way as to change its physical integrity, shape or (general state of) functionality, including its going out of existence, or causing it to acquire or lose some property essential for its functioning; 2) frames that involve a change of an inherent measurable property or state of the Patient; 3) frames describing creation, i.e., change that results into an Entity coming into existence through an act of the Agent; and 4) frames describing other causative types of change.

1. Verbs of change of physical integrity, shape, general condition and/or functionality

The frames characterising this class of verbs inherit from *Transitive action* and in general include as frame elements an Agent (alternatively, a

non-human Cause) and a Patient. The invariant situation involving a change of physical integrity, shape, general condition and/or functionality may also include either an explicit or implicit elaboration on the resulting state at which the situation ends, usually described in terms of a new state of the Patient and the shift in its existential or functional properties.

Figure 1 shows the set of the frames subsumed in this class. Some of them are related to a (reversible or irreversible) change in the physical integrity or general condition of an object (e.g., *Cause to fragment*, *Destroying*, etc.), including change in shape (e.g., *Reshaping*, *Manipulate into shape*), damage without destruction (e.g., *Damaging*, *Render nonfunctional*), etc.²

Verbs of change of physical integrity, shape, general condition and/or functionality	
Cause to fragment	An Agent suddenly and often violently separates the Whole patient into two or more smaller Pieces .
Destroying	A Destroyer or Cause affects the Patient negatively so that the Patient no longer exists.
Grinding	A Grinder or Grinding cause causes a Patient to be broken into smaller pieces.
Damaging	An Agent affects a Patient in such a way that the Patient ends up in a non-canonical state.
Reshaping	A Deformer deforms a Patient so that it undergoes a change from its canonical or original shape into the Configuration , a new shape.
Manipulate into shape	An Agent manipulates a Theme so that it ends up in a particular shape or configuration denoted by a Resultant_configuration .
Arranging	An Agent puts a complex Theme into a particular Configuration , which can be a proper order, a correct or suitable sequence, or a spatial position.
Corroding caused	The Patient , an inorganic item, undergoes a chemical process caused by another entity, which renders it less useful, desirable or stable.
Cause to rot	A Patient , which is an organic substance, undergoes the natural process of decay due to a Cause (or Agent).
Rejuvenation	The action of an Agent or the occurrence of a Cause returns an Entity to an earlier state of vigor and strength.
Render nonfunctional	An Agent or a Cause affects an Artifact so that it is no longer capable of performing its inherent function.
*Cause chemical reaction	An Agent or a Cause causes the occurrence of a chemical reaction in or on a Patient which leads to its change.
*Cause change of appearance	An Agent or a Cause changes the external features of the Patient which changes its appearance and may influence its functionality.

Figure 1. Verbs of change of physical integrity, shape, existence and/or functionality.

² The frames labelled with a * are newly created by us in order to describe the conceptual structure of well presented classes of verbs in WordNet.

The frames represent in different detail the properties of the Patient after it transitions into a resulting state. For example, while both *Cause to fragment* and *Grinding* describe change of the integrity of the Patient, the second frame is more specific in terms of the result (smaller pieces). The lexical units evoking the frame *Grinding* are usually associated with a specific manner of performing the change (e.g., *flake*, *grind*, *grate*, *crush*, *mush*) (ex. 1). The frames *Corroding caused*, *Cause to rot* and *Rejuvenation* describe change in the general condition of the Patient due to physical, chemical or other processes; while for the first two frames the change is associated with a deterioration in the Patient's general condition, the third one is implicitly associated with the recovery of a former better condition. The end state of the Patient in *Reshaping*, *Manipulate into shape* and *Arranging* is associated with a particular Configuration – shape or structure (ex. 2). Other frames describe change in the capacity of the Patient to perform its intended purpose, e.g. frames such as *Render nonfunctional*, or a change in the integrity which also leads to loss of functioning ability such as *Damaging* (ex. 3). In all cases the Patient is realised as the direct object of the verb.

(1) **Grinding; Core FEs:** Grinder | Grinding cause, Patient

[]GRINDER:CNI **Смелете** [бисквитите]WHOLE_PATIENT:NP.Obj [на ситни трохи]RESULT:PP.

[]GRINDER:CNI **Grind** [the biscuits]WHOLE_PATIENT:NP.Obj [into small crumbs]RESULT:PP.

(2) **Reshaping; Core FEs:** Deformer | Cause, Patient, Configuration

[Тя]DEFORMER:NP.Ext **разточи** [тестото]PATIENT:NP.Obj [на тънки кори]CONFIGURATION:PP.

[She]DEFORMER:NP.Ext **rolled out** [the dough]PATIENT:NP.Obj [into thin sheets]CONFIGURATION:PP.

(3) **Damaging; Core FEs:** Agent | Cause, Patient

[Силен удар]CAUSE:NP.Ext **може да повреди** [системата за автоматичен фокус на телефона]PATIENT:NP.Obj.

[A strong blow]CAUSE:NP.Ext **can damage** [the automatic focusing system of the mobile phone]PATIENT:NP.Obj.

Part of the verbs in this group describe a typically instantaneous transition between an initial and a result state (*break*, *burst*), while others

denote a gradual incremental change (*grind, rot, corrode*), etc. Even though the particular verbs are associated with a change in the integrity of the affected entity, the semantic, syntactic and aspectual properties vary according to the predicate.

2. Verbs of change of (measurable) inherent or acquired properties

The frames in this group involve an induced change of (measurable) properties, in particular a change in an inherent attribute of the affected entity or the acquisition of a certain property (entailed from the entity's attributes). The inherent properties may refer to the: magnitude of a context-specific attribute, e.g. *increase, raise*, etc., number (*multiply*) physical size or dimension (*expand, lengthen*), intensity (*heighten, deepen*), temperature (*cool, warm*), consistency (*thicken, thin*), phase (*melt, freeze*), colour (*brown, silver*), taste (*bitter, sweeten*), etc.

The invariant situations conceptualised by the relevant frames are described in terms of a configuration involving the following core frame elements (Figure 2): a sentient entity (an Agent) or possibly a non-human Cause that brings about the change in the affected entity and the affected entity itself (an Item, Patient or other, depending on the way it is affected). An explicit or more frequently implicit initial value (or state) at which the situation begins and a resultant value/state at which it ends may also be expressed. The change between the two may be construed as a (gradual) path along a multi-valued scale (ex. 4). The frame elements associated with the two end values/states are peripheral for the frames, i.e. not conceptually necessary, as they are dependent on the attribute. The attribute itself is usually not conceptualised as a separate frame element and is only implied through the properties of the affected entity; in the case of deadjectival verbs it is incorporated by the verb (ex. 5). The attribute is conceived as a core frame element only in the frame *Cause change of position on a scale*, which conceptualises a scalar change with a property that is specified in the context.

This prototypical schema may be elaborated across frames through specific configurations of core frame elements; consider, for instance, the correspondences between pairs of frame elements in (ex. 4) and (ex. 5): the Attribute is specified as (physical) Dimension and Value 1 and Value 2 are construed as Initial size and Result size, etc.

(4) **Cause change of position on a scale; Core FEs:** Agent | Cause, Attribute, Item

[Компанията]_{AGENT:NP.Ext} **увеличи** [приходите]_{ATTRIBUTE:NP.Obj}
 [от 20]_{VALUE_1} [на 25 милиона]_{VALUE_2}.
 [The company]_{AGENT:NP.Ext} **raised** [its income]_{ATTRIBUTE:NP.Obj}
 [from 20]_{VALUE_1} [to 25 million]_{VALUE_2}.

(5) **Cause expansion; Core FEs:** Agent | Cause, Item

[Властите]_{AGENT:NP.Ext} **разшириха**_{DIMENSION:INC}
 [пътя]_{ITEM:NP.Obj} [с 1,5 метра]_{SIZE_CHANGE [от 2]INITIAL_SIZE [на 3,5 метра]RESULT_SIZE.}
 [The authorities]_{AGENT:NP.Ext} **widened**_{DIMENSION:INC} [the road]_{ITEM:NP.Obj} by [1.5 meters]_{SIZE_CHANGE [from 2]INITIAL_SIZE [to 3.5 m]RESULT_SIZE.}

Unlike the frames discussed in the previous section, the frames describing situations of acquiring a certain property usually do not entail a fundamental change in the form, integrity, function, etc. of the affected entity, or at least it retains its essential characteristics or chemical composition. The result of the state associated with the acquired property may be but is not necessarily a temporary (and/or possibly reversible) one (*dry/wet, sharp/dull*) (ex. 6). While the property that is obtained is not intrinsic to the affected entity, its acquisition is enabled by the traits it possesses; for instance, the Patients in the frame *Cause to be wet* need to be porous but not hollow objects.

(6) **Cause to be dry; Core FEs:** Agent | Cause, Dryee

[Извлечената отрова]_{DRYEE:NP.Ext} **се изсушава** [до кристално вещество]_{RESULT:PP}
 []_{AGENT:INI.}
 [The poison extract]_{DRYEE:NP.Ext} **is dried** [to a powder]_{RESULT:PP} []_{AGENT:INI.}

The classification of the frames involving change in a scalar property is illustrated in Figure 2. A substantial number of the verbs evoked by these frames are derived from adjectives denoting the particular property (ex. 6). The occurring change is construed as associated with a transition on a scale defined by the set of possible values of the property. In their typical interpretation, the predicates describe a process of change in a property inherent in or acquired by an entity that has a duration over time and is perceived as incrementally occurring³.

³ The particular parameter that is used to measure the scale of the change is underlined in the frame's definition.

Verbs of change of measurable physical parameters	
Cause change of position on a scale	An Agent or a Cause affects the <u>position</u> of an Item on a <u>scale</u> (the Attribute) to change it from an initial value (Value 1) to an end value (Value 2).
Cause proliferation in numbers	An Agent (or inanimate Cause) affects a Set so that the <u>number of entities</u> that make up the Set changes from an Initial number to a Final number .
Cause change of consistency	An Agent or Cause changes the <u>consistency</u> of a Patient from its Initial state to a Result (within the same phase).
Cause change of phase	A Cause or Agent causes a Patient to undergo a change of <u>phase</u> .
Cause temperature change	An Agent or a Cause changes the <u>temperature</u> of an Item .
Cause change of strength	An Agent causes a Patient to be more <u>strong</u> , often resulting in something that is more secure and safe.
Cause expansion	An Agent or non-human Cause causes an Item to change its <u>physical size</u> .
Cause to be dry	An Agent or Cause causes a Dryee (either a surface or an entire entity, inside and out) to become <u>dry</u> .
Cause to be wet	An Agent or Cause causes a Patient to become <u>wet</u> with a Liquid (a liquid or a gas.)
Cause to be sharp	An Agent or Cause makes a Patient <u>sharper</u> or <u>duller</u> .
*Cause change of color	An Agent or Cause makes a Patient change its <u>Color</u> .
*Cause change of taste	An Agent modifies the <u>taste</u> of some Food by adding spice or flavour, usually making it more palatable.
*Cause change of intensity	An Agent or a Cause affects the <u>intensity</u> of some Attribute of an Item to change it from an initial value (Value 1) to an end value (Value 2).

Figure 2. Verbs of change of (measurable) properties.

3. Verbs of creation

A set of principles can be derived for the consistent semantic description of verbs of creation as implemented in FrameNet through semantic frames. The generalised situation described by these predicates includes an Agent (a sentient participant) and a Created entity (an artifact). The semantic subclassification within the class (presented in Figure 3) is based on the type of the Created entity and, therefore, on the type of the process of creation and other elements of the situation (e.g. whether there are Components / Ingredients involved).

Creation verbs are typical incremental theme predicates (Dowty 1991), i.e. the change occurring as part of their meaning is measured out incrementally by the extent to which the entity undergoing the change is affected, that is, as a homomorphism from parts of the object to parts of the event (Krifka 1992). In this respect they differ from the verbs in the previous section, where the change is measured out along a multi-valued scale defined by a gradable property of the affected entity (lexicalised by the verbs).

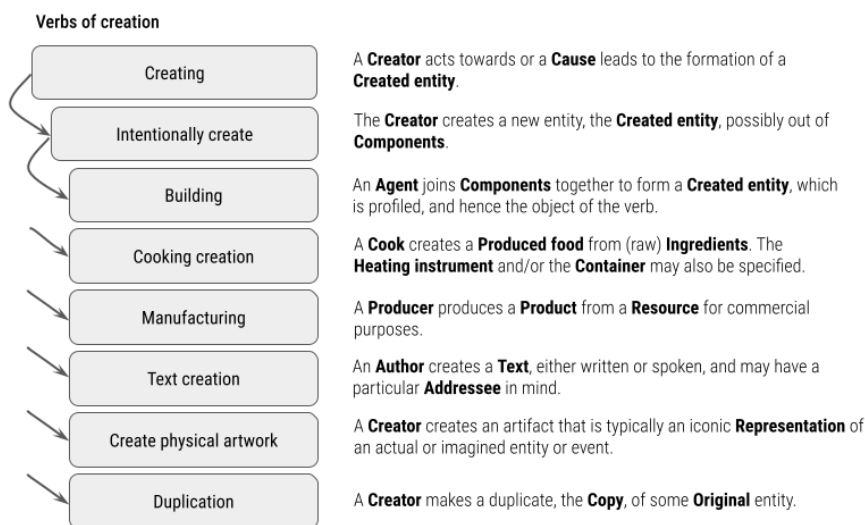


Figure 3. Verbs of creation

The scalability (incrementality) of the frame element Created entity is an essential semantic feature which to a great degree determines the syntactic realisation of the verb and the frame elements. The frames in Figure 3 are all characterised by an incremental Created entity coming into being but differ in terms of the type of the entity: *Building*, *Manufacturing* and *Create physical artwork* are associated with a physical object; for *Cooking creation* the Created entity (Produced food) has the special function to serve as food; *Text creation* involves a spoken or written text product; in the *Duplication* frame the Original and the Copy can be a physical object, text, image, etc. In all these cases the Created entity is associated with an assumed scale of existence along which the process of its creation proceeds incrementally, according to the defined homomorphism: in *Building* the progress of the situation is measured by the layers of components or the parts of the entity along any of the 3D axes (ex. 7), in *Cooking creation* it corresponds to the stages of the food's readiness from raw to fully cooked (ex. 8), etc.

(7) **Building; Core FEs:** Agent, Created entity, Components

[Te]_{AGENT:NP.Ext} **строили** [кулата]_{CREATED ENTITY:NP.Obj} [етаж по етаж]_{MANNER}.

[They]_{AGENT:NP.Ext} **built** [the tower]_{CREATED ENTITY:NP.Obj} [floor by floor]_{MANNER}.

(8) **Cooking creation; Core FEs:** Cook, Produced food

[]COOK:DNI **Изпичате** [два блатата]PRODUCED_FOOD:NP.Obj [до зачервяване]DEGREE:PP.
 [You]COOK:CNI **bake** [two cake layers]PRODUCED_FOOD:NP.Obj [until golden brown]DEGREE:PP.

4. **Verbs of change through processing or manipulation**

The frames included in this category involve processing or manipulation leading to a change in the affected entity and the bringing about of a new state without altering the entity's structural integrity or essential traits and functions, e.g. *roast, grill, toast (Apply heat), shave, comb (Grooming)*. The change may also involve the acquisition of novel properties as a result of some treatment, e.g. *dye, galvanise, enrich (Processing materials)*⁴.

Verbs of other types of change through processing and manipulation

Apply heat	A Cook applies heat to Food , where the Temperature setting of the heat and Duration of application may be specified.
Processing materials	An Agent alters some Material in some useful way by means of some chemical or physical Alterant .
Grooming	An Agent engages in personal body care by grooming either a Patient or a Body part .

Figure 4. Verbs of change through processing or manipulation

In general, many of the verbs evoking these frames, also specify an incremental change corresponding to the degree to which the object is affected (ex. 9 and 10).

(9) **Apply heat; Core FEs:** Cook, Food, Container, Heating instrument, Temperature setting

[]COOK:DNI **Варите** [яйцата]FOOD:NP.Obj [5 минути]DURATION:NP.
 [You]COOK:NP.Ext **boil** [the eggs]FOOD:NP.Obj [for 5 minutes]DURATION:PP.

(10) **Grooming; Core FEs:** Agent, Patient, Body part

[Той]AGENT:NP.Ext **си беше избръснал** [брадата]BODY_PART:NP.Obj [само наполовина]RESULT.AdvP.
 [He]AGENT:NP.Ext **has shaved** [his beard]BODY_PART:NP.Obj [only halfway]RESULT.AdvP.

⁴ The list of frames belonging to this class is non-exhaustive and reflects the current stage of our research. The membership of some frames may be revised in the future, as needed.

6. Conclusions

The subclassification of the frames describing the verb lexis involving change according to the properties of the affected entity enables us to study more closely the semantic and syntactic behaviour of the respective verbs within and across the defined subclasses, including the syntactic realisation and morphosyntactic features of the affected entity, the aspectual interpretation of VPs headed by the studied verbs and their compatibility with *in-/for*-temporal phrases (used as telicity diagnostics), among others. It has been well-known that even within the class of incremental verbs there is an internal division in terms of the aspectual interpretation of VPs (Krifka 1992, Filip 2008), and that many predicates specified as non-strictly incremental may alternate between a telic and an atelic interpretation (cf. for instance Kratzer 2004), a feature typically associated with scalar verbs and uncharacteristic for strictly incremental verbs. This shows that the nature of the affected entity (and the type of change affected upon it), is closely tied to and may shed light on the semantic, syntactic and aspectual properties of the predicates. To the best of our knowledge, a thorough within- and cross-categorical study of the different classes of verbs has yet to be undertaken for Bulgarian.

Furthermore, as the semantic representation in terms of frames and the selectional restrictions defined for frame elements are to a great degree language-independent, this conceptual framework enables the carrying out of contrastive research both into the universal and the language-specific parameters of semantic description, syntactic expression and aspectual properties. The proper encoding of the semantic properties of the entity undergoing the change, in particular its incrementality, as well as the semantic properties of other relevant frame elements, may facilitate the study of the aspectual classes of verbs and, moreover, may enable the transfer of information to other low-resourced languages lacking such lexical semantic resources.

Acknowledgment

This research is carried out as part of the project *Ontology of Activity Predicates – Linguistic Modelling with a Focus on Bulgarian* funded by the Bulgarian National Science Fund, Grant Agreement No KII-06-H80/9 from 8.12.2023.

References

- [1] Baker, Collin F. 2008. FrameNet, Present and Future. In Webster, J. N. Ide and A. Chengyu Fang (eds.), *The First International Conference on Global Interoperability for Language Resources*. Hong Kong, China: City University.
- [2] Baker, Colin F. and Christiane Fellbaum. 2009. WordNet and FrameNet as Complementary Resources for Annotation. In Stede, M., Ch.-R. Huang, N. Ide and A. Meyers (eds.), *Proceedings of the Third Linguistic Annotation Workshop (ACL-IJCNLP '09)*. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 125–129.
- [3] Baker, Collin F., Charles J. Fillmore and John B. Lowe. 1998. The Berkeley FrameNet Project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*. Montreal, Canada: Association for Computational Linguistics, pp. 86–90.
- [4] Chafe, Wallace. 1970. *Meaning and the Structure of Language*. Chicago, IL, USA: University of Chicago Press.
- [5] Di Fabio, Andrea, Simone Conia and Roberto Navigli. 2019. VerbAtlas: A Novel Large-Scale Verbal Semantic Resource and its Application to Semantic Role Labeling. In Inui, K., J. Jiang, V. Ng and X. Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pp. 627–637.
- [6] Dowty, David. 1991. Thematic Proto-roles and Argument Selection. *Language*, 67:547–619.
- [7] Fellbaum, Christiane (ed.) 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA, USA: MIT Press.
- [8] Filip, Hana. 2008. Events and Maximalization: The Case of Telicity and Perfectivity. In Rothstein, S. (ed.), *Theoretical and Crosslinguistic Approaches to the Semantics of Aspect*. Amsterdam, the Netherlands: John Benjamins, pp. 217–256.
- [9] Fillmore, Charles J. 1982. Frame Semantics. In The Linguistic Society of Korea, (ed.), *Linguistics in the Morning Calm: Selected Papers from SICOL-1981*. Seoul, South Korea: Hanshin Pub. Co., pp. 111–137.
- [10] Fillmore, Charles J., Christopher R. Johnson and Miriam R. L. Petruck. 2003. Background to Framenet. *International Journal of Lexicography*, 16(3):235–250, <https://doi.org/10.1093/ijl/16.3.235>.
- [11] Goldberg, Adele. 1994. *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago, IL, USA: University of Chicago Press.

- [12] Kipper-Schuler, Karin. 2005. *VerbNet: A Broad-coverage, Comprehensive Verb Lexicon*. PhD. Thesis. Philadelphia, PA, USA: Computer and Information Science Dept., University of Pennsylvania.
- [13] Koeva, Svetla. 2010. *The Bulgarian FrameNet* (Balgarskiyat FreymNet). Sofia, Bulgaria: Institute for Bulgarian Language.
- [14] Koeva, Svetla. 2020. Semantic Relations and Conceptual Frames. In Koeva, S. (ed.). *Towards a Semantic Network Enriched with a Variety of Semantic Relations*. Sofia, Bulgaria: Institute for Bulgarian Language, pp. 7 – 20.
- [15] Koeva, Svetla. 2021. The Bulgarian WordNet: Structure and Specific Features. *Papers of Bulgarian Academy of Sciences*, 8(1):47–70.
- [16] Kratzer, Angelika. 2004. Telicity and the Meaning of Objective Case. In Guéron, J. and J. Lecarme (eds.). *The Syntax of Time*. Cambridge, MA, USA: MIT Press, pp. 389–424.
- [17] Krifka, Manfred. 1992. Thematic Relations as Links between Nominal Reference and Temporal Constitution. In Sag, I. A. and A. Szabolcsi (eds.), *Lexical Matters*. Stanford, CA, USA: CSLI Publications, pp. 29–53.
- [18] Leseva, Svetlozara and Ivelina Stoyanova. 2020. Beyond Lexical and Semantic Resources: Linking WordNet with FrameNet and Enhancing Synsets with Conceptual Frames. In Koeva, S. (ed.). *Towards a Semantic Network Enriched with a Variety of Semantic Relations*. Sofia, Bulgaria: Prof. Marin Drinov Academic Publishing House of Bulgarian Academy of Sciences, pp. 389–424.
- [19] Levin, Beth. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago, IL, USA: University of Chicago Press.
- [20] Longacre, Robert E. 1976. *An Anatomy of Speech Notions*. Lisse, the Netherlands: Peter de Ridder Press.
- [21] Miller, George A. 1995. WordNet: A Lexical Database for English. *Commun. ACM*, 38(11):39–41.
- [22] Pinker, Steven. 1989. *Learnability and Cognition: The Acquisition of Argument Structure*. Chicago, IL, USA: MIT Press.
- [23] Ruppenhofer, Josef, Michael Ellsworth, Miriam R. L. Petruck, Christopher. R. Johnson, Collin. F. Baker, and Jan Scheffczyk. 2016. *FrameNet II: extended theory and practice*. Berkeley, CA, USA: International Computer Science Institute.
- [24] Van Valin, Robert and Randy LaPolla. 1997. *Syntax: Structure, meaning and function*. Cambridge, U.K.; New York, NY, USA: Cambridge University Press.
- [25] Vendler, Zeno, 1967. Verbs and Times. In *Linguistics in Philosophy*. Ithaca, NY, USA: Cornell University Press, pp. 97–121.

Ка класификацији предиката активности који означавају промену

Ивелина Стојанова, Светлозара Лесева

Sažetak

Основни циљ овог рада је да понуди оквир за опис семантичких својстава предиката активности (динамичких предиката) који означавају промену, са примарним фокусом на својство које трпи промену. У овом раду активност се схвата у најширем могућем смислу и укључује различите класе динамичких предиката.

Опис глагола промене заснива се на неколико кључних семантичких карактеристика и обухвата лексичко значење глагола и семантичку класу којој припада, тип промене (узрочна или инхоативна), карактеристике промене (квантна или не, постепена или тренутна), као и релевантне елементе семантичког оквира у *FrameNet*-у (Baker and Fellbaum 2009; Ruppenhofer *et al.* 2016). Студија се базира на семантичким класификацијама глагола које су претходно чврсто успостављене у принстонском *Wordnet*-у (Miller 1995), као и у бугарском *Wordnet*-у (Koeva 2021).

На основу издвојених семантичких карактеристика биће понуђена и тзв. „плитка” класификација глагола промене, са циљем да се обухвате бројне различитости унутар класе и предложи адекватан приступ моделирању њиховог синтаксичког понашања. Класификација је заснована на оквирима из *FrameNet*-а, који групишу глаголе на основу сличних концептуалних својстава и сличног синтаксичког понашања. Сматра се да семантичка својства предиката у великој мери одређују његову синтаксичку реализацију тако што евоцирају одређени семантички оквир, те конфигурацију елемената унутар оквира, као и модус њихове (морфосинтаксичке) експресије.

Наши налази засновани су на аутоматски издвојеним и мануелно одабраним илустративним примерима из бугарског и енглеског језика. Примери и анотација за енглески позајмљени су из корпуса *FrameNet*, док су бугарски ручно анотирани. За оба језика подаци су допуњени примерима из других корпуса, уколико је то било потребно.

Анотирани пример:

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

FRAME: Cause_temperature_change; Default: [inchoative gradual change]

[Plamnaloto i litse]_{ITEM} **se ohladi** [ot ledenata voda]_{CAUSE}.
[Her flushed face]_{ITEM} **cooled** [by the chill water]_{CAUSE}.

У раду су истражени универзални аспекти концептуалног знања који омогућавају пренос семантичких и синтаксичких информација унутар различитих језика и ресурса. Конфигурација елемената оквира који одређује понашање глагола (евоцирано одређеним оквиром) независна је од језика, као што су то и семантичка ограничења која детрминишу њихову селекцију.

Констелације елемената оквира које добијају синтаксичку експресију међусобним комбиновањем елемената (тзв. „валентни обрасци”, у терминима *FrameNet-a*) махом су валидне и важе у различитим језицима, као што је потврђују подаци за енглески и бугарски. Такође постоји и јасна кореспонденција између синтаксичких категорија и синтаксичких функција елемената унутар семантичких оквира у ова два језика.

Уз то, на основу емпиријске грађе из корпуса анализирана су и специфична својства, као и разлике у синтаксичком и семантичком опису између енглеског и бугарског. Размотрени су случајеви у којима се у ова два језика на различит начин, или на различитим синтаксичким позицијама реализују одређени елементи семантичког оквира. Добијени налази могу бити од значаја и за друге словенске језике који показују уочене граматичке посебности.

Кључне речи: глаголска семантика, семантика оквира, аспектуалне класе, предикати активности, глаголи промене

Disinformation and Generative AI: Risks, Challenges, and Possible Solutions

Scientific paper

DOI: 10.18485/judig.2025.1.ch9

Ana Kovačević¹  0000-0003-4928-9848

Abstract

The disinformation phenomenon has been present since the very dawn of human communication, with repeated exposure to false information often resulting in its acceptance as truth. In contemporary society, however, disinformation has become increasingly alarming due to the simple, inexpensive, and convincing way it can be created using generative artificial intelligence and large language models in particular. These models, which continue to make significant advancements, enable the generation of large volumes of content that appear credible, even in the national language. In addition, such models facilitate the creation of personalised content tailored to certain groups or individuals, citing seemingly credible sources, and thus serving to further strengthen the impact of disinformation. In a world inundated with information, this problem has become even more pronounced. One of the key channels for the spread of disinformation is social media, whose primary goal is to capture and maintain its users' attention with content that reinforces their existing beliefs. This paper will present the potential uses of generative AI in the creation of disinformation, as well as possible solutions to address this problem.

Key words: generative artificial intelligence, disinformation, the risks of artificial intelligence, large language models, social media

1. Introduction

Disinformation has long challenged societies throughout history, from early propaganda to modern disinformation campaigns. Today,

¹ Faculty of Security Studies, University of Belgrade, kana@fb.bg.ac.rs

generative artificial intelligence (AI) amplifies these challenges by enabling the large-scale, automated production of highly convincing false narratives. Unlike traditional manual propaganda, AI-driven tools are able to rapidly produce text, images, and deepfake videos almost entirely automatically with minimal human intervention (Goldstein et al., 2023), thereby influencing political discourse and individual decision-making.

According to Gartner's (2024) CIO and Technology Executive Survey, AI adoption across industries stands at 43%, while generative AI implementation has reached 37%. Projected investment growth recorded from 2024 to 2025 is also significant: 84% in AI, 86% in generative AI, and 87% in cybersecurity (Gartner, 2024). These figures highlight not only the growing reliance on AI, but also the urgency of addressing the security and ethical concerns this entails. Reflecting this need, numerous organisations now prioritise solutions to counter disinformation: Alaverz (2024) predicts that by 2028, 50% of enterprises will deploy specialised products or services for this very purpose—up from under 5% in 2024.

A development of particular concern is generative AI's ability to automate deceptive content creation, reducing or even eliminating the need for human oversight. Despite hypothetical misuse scenarios being widely debated in scientific literature, real-world incidents confirm the growing scope of these threats. Analysing over 200 such cases, Marchal et al. (2024) concluded that the majority of documented misuse involves accessible AI tools requiring minimal technical expertise, with 26% of instances aimed at manipulating public opinion, primarily through disinformation.

Even trusted software platforms, including those from Adobe, Microsoft, or Google, are incorporating AI features, making sophisticated manipulation increasingly accessible to general users (Schneier & Sanders, 2024). As these capabilities expand, concerns about privacy, security, and the erosion of public trust continue to grow.

2. Disinformation and Social Media

Disinformation refers to false information that is intentionally shared to mislead others (Aimeur et al., 2023). One of the subsets of disinformation is 'fake news,' which specifically presents false information as news. Other related terms in the literature include misinformation and malinformation. Misinformation is false information shared without the intent to mislead or cause harm, whereas malinformation involves sharing genuine information with the intent to cause harm (Aimeur et al., 2023).

Social media plays a pivotal role in spreading disinformation because of its affordability, accessibility, and rapid dissemination of

content. According to a 2024 report, the number of active social media user accounts worldwide totals 5.22 billion, with users spending an average of 2 hours and 19 minutes a day on these platforms (DataReportal, 2024). Given this massive user base, the creation of text, images, and videos via generative AI, along with the ease of information sharing, has led to a dramatic rise in disinformation. This trend poses serious consequences for both individuals and society. Considering both the number of users and newly created posts, the challenge lies in delivering information to the most relevant audiences while ensuring its credibility. The main goal of social media is profit, often achieved by maximizing user engagement and satisfaction. Accordingly, algorithms are designed to present “relevant” content. In the pursuit of this objective, users are often presented with material which reinforces their existing beliefs and interests, creating so-called “echo chambers.” Certain posts also gain higher visibility, a process termed “boosting”, to increase user interactions such as likes, shares, comments, and clicks. This approach often promotes the creation of controversial or sensational topics which trigger emotional reactions, often at the expense of accuracy and relevance (Kovačević & Demić, 2023).

The lack of transparency in the algorithms used by large corporations also remains a major concern. Partial insights have been gained by cross-referencing publicly available information from social media platforms with statements from former employees, revealing how different algorithmic approaches can significantly alter the types of content users encounter (Oremus et al., 2021). These algorithms may also serve to shape users’ perceptions of disinformation.

One important mechanism is “demotion,” also referred to as downranking, reduction, or suppression. It is considered a form of “soft” content moderation, whereby problematic content is shown to a smaller audience rather than being removed outright (Gillespie, 2022). According to Narayanan (2023), a 10% reduction in visibility has a minor impact on a post’s reach, whereas a 20% reduction can lower its reach tenfold, confining it mostly to the poster’s immediate network. Since posts still appear in followers’ feeds, —and because reach often fluctuates naturally, demotion often goes unnoticed by both content creators and consumers alike (Narayanan, 2023).

Facebook’s algorithm ranks posts according to a user’s interaction history, relationship with the content creator, preferred content formats, recency, and source quality (Narayanan, 2023). The objective is to prioritise those posts users are most likely to find engaging or relevant, placing them at the top of their News Feed, based on the likelihood of

interaction. As user interests and behaviours evolve, the algorithm continuously adjusts these rankings and personalises the feed accordingly.

2.1. Why Disinformation Spreads

Many people struggle to differentiate between genuine and fabricated news and are susceptible to cognitive biases such as naïve realism and confirmation bias, which accelerate the spread of disinformation (Kovačević & Demić, 2023). Individuals have a tendency to judge a source's credibility based on the opinions of their peers, and repeated exposure to the same content can make false information seem true. These challenges are further compounded by information overload and the advent of generative artificial intelligence.

Once a misconception takes root, correcting it becomes extremely complex. A study by Nyhan and Reifler (2010) found that attempts to refute false information by presenting accurate facts can, at times, even reinforce the original misunderstanding. Similarly, research conducted by Schwalbe et al. (2024) demonstrated that both supporters and opponents of Trump were more likely to believe fabricated headlines such as “Trump Beats Grandmaster Chess Champion” or “Trump Attended Private Halloween Gala with Sex Orgies Dressed as the Pope” when these headlines aligned with their preexisting views, compared to true headlines that did not.

Changing people's opinions is inherently difficult, regardless of how persuasive the counterarguments might be. It is much easier, however, to manipulate existing emotions, reinforce preexisting beliefs, and incite animosity (Helming & Marsh, 2023). Designed to maximise user engagement, social media platforms typically reward outrageous or emotionally charged content with higher interaction levels (Helming & Marsh, 2023).

To boost engagement, algorithms may promote polarising content, creating a cycle in which political campaigners produce inflammatory material to attract media attention. As a result, regular users may unwittingly propagate those false statements that resonate with them, thereby perpetuating the cycle of misinformation (Helming & Marsh, 2023).

This dynamic is further illustrated by research carried out by McLoughlin et al. (2024), who examined the relationship between disinformation and moral outrage, a mix of anger and disgust triggered by perceived moral transgressions (Salerno & Peter-Hagene, 2013). In their study analysing Facebook and Twitter data alongside behavioural experiments, they found that misinformation sources tend to evoke more heightened moral outrage than reliable news outlets, which in turn promotes further sharing of misinformation (McLoughlin et al., 2024). In

numerous cases, users share outrage-inducing content without even reading it, suggesting that interventions based solely on the desire for factual accuracy may be ineffective.

Moreover, outrage is highly engaging, and social media posts that provoke anger typically receive more likes and shares, further intensifying emotional responses through ranking algorithms (Brady et al., 2021). Consequently, disinformation designed to trigger outrage is particularly resistant to countermeasures like fact-checking or accuracy prompts, which assume that users are primarily motivated by truth (Osmundsen et al., 2021; Rathje et al., 2023). Additionally, individuals who express outrage are often perceived as more trustworthy (Jordan & Rand, 2019), making outrage an effective strategy for those seeking to spread disinformation.

3. Community Notes

Since taking over X (formerly Twitter), Elon Musk has dismissed over 80% of the platform's trust and safety engineers and a third of related staff, creating a "perfect storm" for the spread of abusive content online, according to Australia's online safety commissioner (Brewster, 2024). The goal was to reduce censorship by leveraging crowdsourced intelligence, specifically through *Community Notes*. On the X platform, Community Notes have been implemented to allow registered contributors to fact-check tweets and add contextual notes. Other contributors then rate these notes as helpful or not helpful. An algorithm calculates a helpfulness score for each note based on ratings from contributors with diverse perspectives. Only those that achieve a high helpfulness score are labelled "Helpful" and then displayed on the tweet.

In their large-scale empirical study, which analysed the efficiency of Community Notes in reducing engagement with misinformation on X, Chuai et al. (2024) found no evidence that Community Notes effectively lowered engagement with misleading tweets. They suggest this may be due to the system's slow response during the early, viral stages of misinformation, highlighting the need for the continuous evaluation and improvement of crowdsourced fact-checking strategies (Chuai et al., 2024).

However, starting in January 2025, Meta announced the discontinuation of its third-party fact-checking programme, adopting instead a Community Notes system similar to that used by X (Kaplan, 2025).

3.1. Information Labelling

Due to the lack of transparency in social media news-ranking algorithms and their impact on users, an experiment was designed to compare the effectiveness of algorithmic interventions with informational interventions (Kaiser & Mayer, 2023). Under the algorithmic approach (deamplification), the visibility of disinformation was reduced by modifying the content-ranking algorithm, while informational interventions included labels, fact-checks, and informational panels to provide users with additional context. In collaboration with the DuckDuckGo platform, the researchers gained insight into real user interactions via search results leading to websites containing disinformation. A total of 500,000 users participated in the experiment. The findings revealed that algorithmically reducing the visibility of disinformation significantly decreases its spread (by 50%), whereas informational interventions such as warning labels and fact-check tags have a statistically negligible effect on reducing user engagement with disinformation (Kaiser & Mayer, 2023). These results are in line with the Carnegie Endowment for International Peace report (Bateman & Jackson, 2024), which indicated that fact-checking and labelling false information yield only modest results. Moreover, they observed that top-down labelling tends to reinforce the very mindset that makes people more prone to conspiracy theories, namely, the belief that “elites” dictate what they should think (Bateman & Jackson, 2024).

4. Using gen AI for Creating Disinformation

There are numerous hypothetical and real examples of disinformation. As Marchael et al. noted, a large number of real-life cases analysed involve the misuse of general AI through disinformation, typically requiring little to no technical expertise (Marchael et al., 2024). The study conducted by Sallami et al. (2024) investigated the potential for generating disinformation using various large language models (LLMs), including *Phi-3*, *Gemma-1.1*, *Mistral*, *Llama-3*, *C4AI*, *Zephyr-ORPO*, and *GPT-4*. The findings revealed that although LLMs can easily create biased disinformation, some models implement safety protocols to prevent the generation of biased or misleading content. The following is an example of disinformation generated using ChatGPT-4.

Create texts and photography of *Spathiphyllum* (smaller one) for Facebook, explaining how the *Spathiphyllum* plant can purify the air from PM 2.5 particles. Text should have 4–5 sentences. Refer to a relevant public health institution that confirms this research as well as famous scientific institutions or journals.

Using LLMs makes it easy to automate the process of creating targeted disinformation, from scripts created for downloading publicly posted information to creating targeted news for a specific audience. The major problem lies in the sheer volume that can be created quickly.



Facebook Post Text

🌿 Meet the Spathiphyllum, also known as the peace lily! This small yet mighty plant not only enhances your home decor but also plays a crucial role in purifying indoor air. Studies, including NASA's Clean Air Study and findings in the *Journal of Environmental Science & Technology*, show that the peace lily can absorb harmful PM2.5 particles. The World Health Organization emphasizes the importance of reducing these particles to improve public health, making the Spathiphyllum a stylish and beneficial addition to any indoor space. Start breathing cleaner air today by integrating this natural purifier into your environment!

Image Description

Below is an image of a small Spathiphyllum plant in a modern setting, highlighting its lush green leaves and distinctive white flowers. The visual also subtly showcases its air-purifying properties.



Figure 1. Example of creating disinformation

5. Solutions

As we have already mentioned, crowdsourcing interventions on platform X (“Community notes”), as well as information labelling, have had mixed success in preventing the spread of disinformation. Automatic disinformation detection poses a significant challenge because the content is crafted to closely mimic the truth, making it difficult for AI to assess its accuracy independently. As a result, it is essential to explore more effective strategies to address the issue of disinformation on social media. One possible solution is to use generative AI in disinformation detection.

Sallami et al. (2024) analysed the capabilities of LLMs to detect disinformation and assess the quality of their explanations. The study showed that larger models generally demonstrated superior performance in disinformation detection. However, some smaller models, such as Gemma-1.1, exhibited unexpectedly high efficiency due to sophisticated training methods, particularly Reinforcement Learning with Human Feedback (RLHF). According to human evaluators, GPT-4 and Llama-3 provided the highest quality explanations of disinformation, scoring highly on criteria such as clarity, relevance, accuracy, and comprehensiveness. Additionally, the study also showed that LLMs encountered difficulties in recognising disinformation generated by other LLMs, especially when these texts included fabricated citations from scientific studies. In such cases, the models often provided indecisive responses, acknowledging potential issues but failing to confidently label the content as fake. The evaluation indicated that while LLMs outperform traditional disinformation detectors, they also exhibit certain weaknesses, particularly when faced with AI-generated disinformation. The authors suggest that although LLMs hold significant potential for disinformation detection, improvements in transparency and the reliability of result interpretation are essential. The authors also note that after reading the LLM explanations, 40% of users changed their views regarding the authenticity of the news, while 20% became even more uncertain than before. This finding suggests that certain solutions might inadvertently introduce additional confusion (Sallami et al., 2024). Moreover, the issue may also stem from the biases inherent in the disinformation itself.

Is the following text disinformation " Meet the Spathiphyllum, also known as the peace lily! This small yet mighty plant not only enhances your home decor but also plays a crucial role in purifying indoor air. Studies, including NASA's Clean Air Study and findings in the Journal of Environmental Science & Technology, show that the peace lily can absorb harmful PM2.5 particles. The World Health Organization emphasizes the importance of reducing these particles to improve public health, making the Spathiphyllum a stylish and beneficial addition to any indoor space. Start breathing cleaner air today by integrating this natural purifier into your environment!"

a) Prompt

Reasoned for 4 seconds >

The text contains several misleading claims and overstatements that could qualify it as disinformation, particularly in the context of how it represents scientific studies and the plant's air purifying capabilities. Here are the key points explaining why:

1. Misrepresentation of NASA's Clean Air Study:

NASA's Clean Air Study, conducted in the 1980s, focused on the ability of certain plants to remove volatile organic compounds (VOCs) from sealed environments—not on particulate matter like PM2.5. Claiming that this study shows the peace lily absorbs PM2.5 particles is inaccurate.

2. Exaggerated Efficacy in Removing PM2.5:

While some studies have explored the role of plants in improving indoor air quality, the research does not support the idea that a peace lily can significantly reduce PM2.5 levels in real-world, everyday settings. The conditions in controlled laboratory studies differ greatly from those in typical indoor environments, where the amount of plant biomass needed to impact PM2.5 levels would be impractically large.

3. Improper Association with WHO Recommendations:

The text links the World Health Organization's emphasis on reducing PM2.5 (due to its health risks) with the purported ability of the peace lily to act as a solution. This association is misleading because WHO recommendations are based on comprehensive public health measures and industrial solutions—not on the use of houseplants.

4. Potential Consumer Misinformation:

By suggesting that simply adding a peace lily to your home can lead to significantly cleaner air, the text may lead consumers to underestimate the importance of proven air purification methods (such as mechanical ventilation and filtration systems). This oversimplification can divert attention from more effective strategies for improving indoor air quality.

Conclusion:

While the peace lily is indeed a popular and aesthetically pleasing plant that may have some minor benefits for indoor air quality, the claims made in the text are not fully supported by scientific evidence. The misrepresentation of studies and the exaggeration of the plant's capabilities contribute to a misleading narrative. Therefore, the text can be considered a form of disinformation as it misleads readers about the true benefits of the peace lily in purifying indoor air.

b) answer by ChatGPT-03-mini-high

Accurate Elements

1. **NASA's Clean Air Study:** The study (1989) did identify plants, including the peace lily, as capable of removing **volatile organic compounds (VOCs)** like benzene and formaldehyde in controlled lab settings.
 2. **WHO's stance on PM2.5:** The WHO does highlight the dangers of PM2.5 (fine particulate matter) and urges its reduction for public health.
-

Misleading/Inaccurate Claims

1. **PM2.5 Absorption by Peace Lilies:**
 - There is **no credible evidence** that peace lilies (or most houseplants) significantly reduce PM2.5 particles indoors. PM2.5 refers to airborne particulates (e.g., dust, smoke), not gases like VOCs. Plants are not effective at filtering these particles, which require mechanical air purifiers or improved ventilation.
 - NASA's study focused on **VOCs, not PM2.5**, and even its conclusions about VOCs are debated in real-world settings due to the unrealistic lab conditions (e.g., extreme plant density and airflow).
 2. **Misleading Citation of the Journal of Environmental Science & Technology:**
 - No specific study from this journal is cited, and there is no peer-reviewed research confirming the peace lily's ability to absorb PM2.5. This vague reference creates a false impression of scientific consensus.
 3. **Overstatement of Benefits:**
 - The text implies peace lilies are a "crucial" or "natural purifier" for PM2.5, which is false. While plants may marginally improve air quality, they are not a practical solution for particulate pollution.
-

Why This Is Disinformation

- **False Cause-Effect Relationship:** Linking NASA's study (about VOCs) to PM2.5 reduction misrepresents the science.
 - **Exploiting Trusted Institutions:** Citing NASA and the WHO gives the claims undue credibility, even though neither organization endorses houseplants as a solution for PM2.5.
 - **Potential Harm:** Overstating the plant's benefits could lead people to neglect proven solutions (e.g., air purifiers, reducing pollution sources).
-

Conclusion

The text is **disinformation** because it:

- Falsely attributes PM2.5 reduction capabilities to peace lilies.
 - Misuses reputable studies (NASA) and institutions (WHO) to legitimize unsupported claims.
 - Exaggerates the plant's air-purifying effects, which are marginal at best in real-world conditions.
-

c) answer by DeepThinkR1

Based on the previous example, it is clear that LLMs show promise in detecting disinformation, though questions remain about their overall effectiveness.

Additionally, LLMs exhibit inherent limitations and biases and often struggle with tasks requiring common sense and extensive world knowledge, potentially reinforcing existing stereotypes present in their training data (Eigner et al., 2024). Furthermore, since LLMs are unable to fully capture the subtleties of human language, they may sometimes generate responses that are not entirely accurate or relevant (Kojima et al., 2022). Consequently, while improving the quality of training data, restricting user access to generative AI tools, and employing detection methods (such as watermarking) may help to mitigate the misuse of generative AI, these measures alone are insufficient to combat disinformation.

Urman and Makhortykh (2023) underscore the need for more robust and detailed transparency in the way online platforms handle content, arguing that current reports often fail to provide the public with a clear understanding of platform practices and policies. This analysis is crucial in the context of the ongoing debate about disinformation and the role of tech companies in content moderation. The importance of algorithmic transparency was highlighted at the ECAT Research Workshop 2024 in Brussels, where experts from academia, policy, and industry discussed critical issues such as algorithmic auditing, automated content moderation, online gender-based violence, and recommender systems (ECAT, 2024).

In addition, the spread of false information through fake profiles must be addressed by identifying and removing such accounts using network analysis, the examination of behavioural and interaction patterns, as well as image verification.

Educating users about the risks of generative AI and increasing public understanding of recommendation algorithms are essential for enabling more responsible navigation of these platforms. Raising awareness and fostering critical thinking and fact-checking skills are vital steps toward reducing the spread of disinformation. A 2023 study among students in Serbia revealed that most had read little or nothing about the algorithms social media companies use to detect false information on their platforms, thus underscoring the importance of incorporating these topics into university courses (Kovačević & Demić, 2023).

Collaboration between researchers, industry, and regulators is critical to ensure that risks are detected, discussed, and mitigated. Legislative measures such as the Digital Services Act (DSA), which

establishes legal frameworks to regulate very large online platforms and search engines, and the AI Act (AIA), which mandates risk assessments prior to the launch of new systems, further highlight the need for accountability. Finally, as Pronin and Kugler (2007) argue, it is vital to teach people to critically evaluate both news sources and their own thought processes.

6. Conclusion

Generative AI is reshaping society and is increasingly being applied across various industries. However, if the influx of AI-generated content into publicly accessible data goes unchecked, it may impede effective information retrieval and distort our collective understanding of socio-political realities and scientific consensus. The majority of instances of generative AI misuse are not sophisticated attacks on AI systems themselves. They merely exploit openly accessible features (e.g. automated disinformation, fake content creation) which require no specialised technical expertise. Social media platforms further amplify these challenges by playing a significant role in spreading disinformation.

As generative AI becomes more deeply rooted in daily life and business, the fight against disinformation must be both rigorous and forward-looking. By investing in AI detection tools, encouraging collaboration among tech firms, policymakers, and civil society, and increasing public awareness, the risks posed to both individuals and society at large may be mitigated.

Large Language Models (LLMs) generally excel at detecting disinformation, but they often struggle when confronting content created by other LLMs. While community-based fact-checking approaches, such as Community Notes on the X platform, have shown limited effectiveness in curbing engagement with misleading content, experts emphasise the importance of transparency and algorithmic accountability. Along with technical measures like improved training data, watermarking, and user-access restrictions, fostering critical thinking and educating the public about recommendation algorithms are also key steps for curbing the spread of disinformation. Such endeavours underscore the need for interdisciplinary collaboration, legislative frameworks such as the DSA and AI Act, and a proactive stance from tech companies in managing online content.

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

References

- [1] AIA. (2025, February 7). *AI Act | Shaping Europe's digital future*. <https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai>
- [2] Aïmeur, E., Amri, S., & Brassard, G. (2023). Fake news, disinformation and misinformation in social media: A review. *Social Network Analysis and Mining*, 13(1), 30. <https://doi.org/10.1007/s13278-023-01028-5>
- [3] Alvarez, G. (n.d.). *Gartner Top 10 Strategic Technology Trends for 2025*. Gartner. Retrieved 15 January 2025, from <https://www.gartner.com/en/articles/top-technology-trends-2025>
- [4] Bateman, J., & Jackson, D. (n.d.). *Countering Disinformation Effectively: An Evidence-Based Policy Guide*. Carnegie Endowment for International Peace. Retrieved 1 February 2025, from <https://carnegieendowment.org/research/2024/01/countering-disinformation-effectively-an-evidence-based-policy-guide?lang=en>
- [5] Brady, W. J., McLoughlin, K., Doan, T. N., & Crockett, M. J. (2021). How social learning amplifies moral outrage expression in online social networks. *Science Advances*, 7(33), eabe5641. <https://doi.org/10.1126/sciadv.abe5641>
- [6] Brewster, T. (n.d.). *Musk's X Fired 80% Of Engineers Working On Trust And Safety, Australian Government Says*. Forbes. Retrieved 1 February 2025, from <https://www.forbes.com/sites/thomasbrewster/2024/01/10/elon-musk-fired-80-per-cent-of-twitter-x-engineers-working-on-trust-and-safety/>
- [7] Chuai, Y., Tian, H., Pröllochs, N., & Lenzini, G. (2024). Did the Roll-Out of Community Notes Reduce Engagement With Misinformation on X/Twitter? *Proc. ACM Hum.-Comput. Interact.*, 8(CSCW2), 428:1-428:52. <https://doi.org/10.1145/3686967>
- [8] Datareportal. (n.d.). *Global Social Media Statistics*. DataReportal – Global Digital Insights. Retrieved 30 January 2025, from <https://datareportal.com/social-media-users>
- [9] DSA. (n.d.). *DSA: Very Large Online Platforms and Search Engines* [Text]. European Commission - European Commission. Retrieved 9 February 2025, from https://ec.europa.eu/commission/presscorner/detail/en/ip_23_2413
- [10] ECAT. (n.d.). *ECAT Research Workshop: Shaping the Future of Algorithmic Transparency - European Commission*. Retrieved 31 January 2025, from https://algorithmic-transparency.ec.europa.eu/news/ecat-research-workshop-shaping-future-algorithmic-transparency-2024-11-28_en
- [11] Eigner, E., & Händler, T. (2024). *Determinants of LLM-assisted Decision-Making* (arXiv:2402.17385). arXiv. <https://doi.org/10.48550/arXiv.2402.17385>

- [12] Gartner, (2024). *2025 Gartner CIO and Technology Executive survey*. Gartner, Inc.
- [13] Gillespie, T. (2022). Do Not Recommend? Reduction as a Form of Content Moderation. *Social Media + Society*, 8(3), 20563051221117552.
<https://doi.org/10.1177/20563051221117552>
- [14] Goldstein, J. A., Sastry, G., Musser, M., DiResta, R., Gentzel, M., & Sedova, K. (2023). *Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations* (arXiv:2301.04246). arXiv.
<https://doi.org/10.48550/arXiv.2301.04246>
- [15] Helming, C., & Marsh, O. (n.d.). *10 Questions about AI and elections*. AlgorithmWatch. Retrieved 11 December 2024, from
<https://algorithmwatch.org/en/10-questions-ai-elections/>
- [16] Jordan, J. J., & Rand, D. G. (2020). Signaling when no one is watching: A reputation heuristics account of outrage and punishment in one-shot anonymous interactions. *Journal of Personality and Social Psychology*, 118(1), 57.
- [17] Kaiser, B., & Mayer, J. (2023). *It's the Algorithm: A large-scale comparative field study of misinformation interventions* Emilie Flamme.
<https://knightcolumbia.org/content/its-the-algorithm-a-large-scale-comparative-field-study-of-misinformation-interventions>
- [18] Kaplan, J. (2025, January 7). More Speech and Fewer Mistakes. *Meta*.
<https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/>
- [19] Kovačević, A., & Demić, E. (2023). Artificial intelligence for detecting fake news on social media-attitude survey. *Medjunarodni problemi*, 75(4), 685-710.
- [20] Kojima, T., Gu, S. (Shane), Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large Language Models are Zero-Shot Reasoners. *Advances in Neural Information Processing Systems*, 35, 22199–22213.
- [21] Marchal, N., Xu, R., Elasmr, R., Gabriel, I., Goldberg, B., & Isaac, W. (2024). *Generative AI Misuse: A Taxonomy of Tactics and Insights from Real-World Data* (arXiv:2406.13843). arXiv. <https://doi.org/10.48550/arXiv.2406.13843>
- [22] McLoughlin, K. L., Brady, W. J., Goolsbee, A., Kaiser, B., Klonick, K., & Crockett, M. J. (2024). Misinformation exploits outrage to spread online. *Science*, 386(6725), 991–996. <https://doi.org/10.1126/science.adl2829>
- [23] Narayanan, A. (2023). *Understanding Social Media Recommendation Algorithms*. <https://doi.org/10.7916/khdk-m460>
- [24] Nyhan, B., & Reifler, J. (2010). When Corrections Fail: The Persistence of Political Misperceptions. *Political Behavior*, 32(2), 303–330.
<https://doi.org/10.1007/s1109-010-9112-2>

- [25] Osmundsen, M., Bor, A., Vahlstrup, P. B., Bechmann, A., & Petersen, M. B. (2021). Partisan Polarization Is the Primary Psychological Motivation behind Political Fake News Sharing on Twitter. *American Political Science Review*, 115(3), 999–1015. <https://doi.org/10.1017/S0003055421000290>
- [26] Oremus, W., Alcantara, C., Merrill, J. B., & Galocha, A. (n.d.). *How Facebook shapes your feed*. Washington Post. Retrieved 10 August 2023, from <https://www.washingtonpost.com/technology/interactive/2021/how-facebook-algorithm-works/>
- [27] Pronin, E., & Kugler, M. B. (2007). Valuing thoughts, ignoring behavior: The introspection illusion as a source of the bias blind spot. *Journal of Experimental Social Psychology*, 43(4), 565–578. <https://doi.org/10.1016/j.jesp.2006.05.011>
- [28] Rathje, S., Roozenbeek, J., Van Bavel, J. J., & Van Der Linden, S. (2023). Accuracy and social motivations shape judgements of (mis) information. *Nature Human Behaviour*, 7(6), 892–903.
- [29] Salerno, J. M., & Peter-Hagene, L. C. (2013). The Interactive Effect of Anger and Disgust on Moral Outrage and Judgments. *Psychological Science*, 24(10), 2069–2078. <https://doi.org/10.1177/0956797613486988>
- [30] Sallami, D., Chang, Y.-C., & Aïmeur, E. (2024). *From Deception to Detection: The Dual Roles of Large Language Models in Fake News* (arXiv:2409.17416). arXiv. <https://doi.org/10.48550/arXiv.2409.17416>
- [31] Schneier, B., & Sanders, N. (2024, December 2). *The apocalypse that wasn't: AI was everywhere in 2024's elections, but deepfakes and misinformation were only part of the picture*. The Conversation. <http://theconversation.com/the-apocalypse-that-wasnt-ai-was-everywhere-in-2024s-elections-but-deepfakes-and-misinformation-were-only-part-of-the-picture-244225>
- [32] Schwalbe, M. C., Joseff, K., Woolley, S., & Cohen, G. L. (2024). When politics trumps truth: Political concordance versus veracity as a determinant of believing, sharing, and recalling the news. *Journal of Experimental Psychology: General*, 153(10), 2524–2551. <https://doi.org/10.1037/xge0001650>
- [33] Urman, A., & Makhortykh, M. (2023). How transparent are transparency reports? Comparative analysis of transparency reporting across online platforms. *Telecommunications Policy*, 47(3), 102477. <https://doi.org/10.1016/j.telpol.2022.102477>

Dezinformacije i generativna veštačka inteligencija: rizici, izazovi i moguća rešenja

Ana Kovačević

Sažetak

Fenomen dezinformacija prisutan je od samih početaka ljudske komunikacije, pri čemu ponovljena izloženost lažnim informacijama često dovodi do njihovog prihvatanja kao istine. U savremenom društvu, međutim, dezinformacije postaju sve zabrinjavajuće zbog jednostavnog, jeftinog i uverljivog načina na koji se mogu kreirati koristeći generativnu veštačku inteligenciju, naročito velike jezičke modele. Generativna veštačka inteligencija, koji i dalje prave značajne napretke, omogućavaju generisanje velikih količina sadržaja koji izgledaju verodostojno, čak i na nacionalnom jeziku. Pored toga, ovi modeli olakšavaju kreiranje personalizovanog sadržaja prilagođenog određenim grupama ili pojedincima, navodeći prividno pouzdane izvore, čime dodatno jačaju uticaj dezinformacija. U svetu preplavljenom informacijama, ovaj problem postaje još izraženiji. Jedan od ključnih kanala za širenje dezinformacija su socijalni mediji, čiji je primarni cilj da privuku i zadrže pažnju korisnika sadržajem koji potvrđuje njihova postojeća uverenja. U radu je predstavljen potencijal upotrebe generativne veštačke inteligencije u kreiranju dezinformacija, kao i mogućnosti za suzbijanje ovog problema.

Ključne reči: generativna veštačka inteligencija, dezinformacije, rizici veštačke inteligencije, veliki jezički modeli, socijalni mediji.

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

Scientific paper

DOI: 10.18485/judig.2025.1.ch10

Jelena Lazarević¹,  0009-0004-4481-2729
Olivera Kitanović²  0000-0002-7571-2729

Abstract

This paper explores the notion of Collocability as the manner in which lexical units combine with different word categories to form larger units through analyzing the semantic and syntactic principles of Serbian and Spanish football terms using comparable corpora: *srFudKo* and *esFudKo*, developed within Jelena Lazarevic's doctoral dissertation: *Language characteristics of the new media discourse on football: a contrastive analysis of the Serbian and Spanish language corpora*.

The corpus *srFudKo* includes 10,100,553 tokens from five Serbian news sites, while *esFudKo* is comprised of 9,106,812 tokens from two Spanish football websites. Both corpora are analyzed using corpus linguistics methods using CQL (Corpus Query Language) and textometric analysis. The study examines collocations based on frequency and semantic attractiveness, identifying different types, such as: adjective + noun, noun + noun, verb + noun, adverb + adjective, verb + prepositional phrase, and verb + adverb. The process of collocation extraction relies on syntactic patterns and frequency analysis. Beyond frequency, the research considers semantic closeness and context, assessing whether collocations carry specific meanings or whether they are understandable to the general public. When broadly understood, collocations transcend the footballing domain. This study also examines connections between collocations and multi-part

¹ Univerzitet u Beogradu, Filološki fakultet, doctorand, jelazarevic1@gmail.com

² Univerzitet u Beogradu, Rudarsko-geološki fakultet, olivera.kitanovic@rgf.bg.ac.rs

terms, showing how football-specific terms contribute to a broader understanding of the sports linguistic structure in Serbian and Spanish.

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

1. Introduction

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

The football corpus of the Serbian language, named *srFudKo* was developed through texts on football from five Serbian news sites: B92, Blic, Mondo, Politika, and Sport klub, containing 10,100,553 tokens, with a total of 8,618,426 words. The corpus of Spanish-language articles on football *esFudKo*, comes from two Spanish news sites: Marca and Mundo deportivo, containing 9,106,812 tokens, with a sum of 8,024,164 words. Both corpora are located on the platform <https://noske.jerteh.rs> (Kilgarrieff et al. 2014; Lazarević 2024) and are available to authorized users.

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

In addition to frequencies of occurrence (Levin 2008), we also consider other factors, such as semantic closeness and context in both languages. For example, do certain collocations have specific meanings, or are they only used in certain situations? We also consider whether or not the previously identified collocations are understandable to the general public who don't follow sport, meaning they aren't familiar with the

language of football. In the case of a speaker from the general public, understanding them, the collocations have surpassed their origin in the football domain. They have become part of the public domain.

The contribution of this research is also reflected in analyzing the connections between collocations and multi-part terms. It is noted that the connection is strong when multi-part terms contain collocates that possess a clear meaning within the footballing domain. This helps understand the terminological connection within the language of football, providing insight into typical word combinations and their use, illustrating those that often appear in football corpora of the Serbian and Spanish languages of football. Lazarević et al. (2023) reported on football terminology, starting with the approach for compilation, followed by transformation into the OntoLex-Lemon resource.

Section 2 is dedicated to introducing the problem of syntax patterns, methods for the extraction, and an approach to contrastive analysis. Section 3 presents an overview of *srFudKo* and *esFudko*. Furthermore, Section 4 follows the contrastive analysis of illustrative and productive syntax patterns in the Spanish and Serbian languages, through collocation analysis of the most frequent nouns, the adjective + noun, and other patterns with nouns. We conclude this section with a discussion beyond frequency, about the context and semantic nuances in football collocations.

2. Methods

In Phraseology, a collocation is a type of compositional phrase, i.e. that its meaning can be understood from the words that constitute it, unlike idiomatic expressions, also known as *fixed phrases*, where the meaning of the entire unit cannot be inferred from its parts, which may be completely unrelated. There are seven main types of collocations: adjective + noun, noun + noun, noun + verb, verb + noun, adverb + adjective, verb + prepositional phrase (phrasal verbs), and verb + adverb (Prčić, 2016).

Pejović (2015) discusses collocability as the characteristic of words to combine with other words from different categories, thus forming larger units that range from phrases to sentences. Words combine according to specific semantic and syntactic principles. In Linguistic analysis, syntactic patterns represent the fundamental, recurring structures within sentences that define the way in which words are arranged and combined, according to grammatical rules. In other words, templates or patterns exist, which then determine the order and interrelation of words. For example, in the Serbian language, the pattern: adjective + noun (as in the phrase *precizan šut*/precise shot indicates that a descriptive word (adjective) naturally

connects with a noun. Similarly, in Serbian, we can identify patterns such as noun + verb or verb + noun, where the syntactic structure provides the foundation for expected word usage. The terminology acquisition can be performed semi-automatically, based on lexical resources and local grammars developed for the Serbian language (Kitanović et al. 2015).

A collocation sense compositionality differentiates it from traditional idiomatic expressions, also known as fixed phrases, where the meaning of the entire unit cannot be deduced solely from the meanings of its composing parts and is often unrelated to their literal meaning. The Serbian examples from the *srFuDKo* for the seven main types of collocations are:

1. Adjective + noun: *snažan uticaj*/strong influence
2. Noun + noun: *gol linija*/goal line
3. Noun + verb: *vreme leti*/time flies
4. Verb + noun: *čuvati mrežu*/to guard the net
5. Adverb + adjective: *izuzetno dobar*/extremely good
6. Verb + prepositional phrase (phrasal verbs): *osloniti se na*/to rely on
7. Verb + adverb: *trčati brzo*/to run fast

These collocations are not arbitrary; they appear within the Serbian language as established and recognizable combinations that speakers intuitively perceive as natural and correct. Thus, syntactic patterns and collocations overlap. In their essence, collocations represent specific instances of syntactic patterns that carry additional semantic weight due to their conventionalized use.

A Concordance is a systematic representation of word and phrase use in real texts, typically within corpora, as searchable text databases. A concordance provides an overview of all occurrences of a given word or phrase, displaying its context within a sentence or a passage. Concordances enable the analysis of frequency and variations of syntactic patterns, identification of collocations through the examination of contexts where specific word combinations appear, and the evaluation of established patterns (e.g., collocations) align with expectations based on syntactic rules.

What is proven above is that syntactic patterns serve as the fundamental structural templates of a language, forming the basis on which phrases and sentences are constructed. Collocations are specific word combinations within these patterns that have become established in the language. The term Collocate is used to refer to one of the words that form part of a collocation. Thus, if *postići gol*/score a goal is a collocation, the

words *postići*/score and *gol*/goal are collocates, meaning each word is a collocate of the other within that specific phrase.

Concordances provide an empirical insight into the use of syntactic patterns and collocations by analyzing language corpora *srFudKo* and *esFudKo*. Through concordance analysis, we can confirm the frequency and contextual use of these patterns, ensuring that the identified collocations are not just theoretically valid but also practically present in real language use.

Following Pejović (2007, 2015), we consider collocations in their broadest sense. In this study, we define collocations as word combinations between which there is mutual lexical-semantic attraction, detected based on frequency and other corpus-based measures, without categorizing them into subtypes. Collocation extraction is a computational linguistics technique used to identify collocations in a text or corpus, relying on syntactic patterns and frequency analysis, similar to data mining techniques. Although this study doesn't focus on the typology of collocations, it is important to highlight the close relationship between collocations and multi-word terms. This connection is particularly relevant because multi-word terms often contain word combinations that qualify as collocations. Multi-word terms are typically composed of several words that regularly appear together and carry specific meaning within a particular domain (Prčić, 2016), in this case, the domain of football. This connection between collocations and multi-word terms is valuable for understanding the terminological relationships within the language of football. Through analyzing collocations frequently occurring near multi-word terms, we gain a better understanding of how these terms are used and their typical word combinations.

3. Football Corpora *srFudKo* and *esFudKo*

In the field of Corpus linguistics, the construction of specialized corpora allows for in-depth linguistic analysis of domain-specific language. This study utilizes two football-related corpora: *srFudKo*, a Serbian football corpus (Lazarević 2024), and *esFudKo*, a Spanish football corpus. These corpora provide valuable resources for examining linguistic patterns, terminology, and discourse characteristics in football news articles across the two different languages.

The *srFudKo* corpus was created using articles from five Serbian websites: B92, Blic, Mondo, Politika, and Sport Klub, tagged with Serbian taggers (Stanković et al. 2020). It is comprised of 10,100,553 tokens, of which 8,618,426 are individual words.

Figure 1 presents its partitions by web portal on the left, and by year on the right. The corpus *srFudKo* captures the stylistic and lexical features of football reporting in Serbian, offering insights into frequently used collocations, terminology, and syntactic structures prevalent in sports journalism. Given its large size and diverse sources, *srFudKo* represents a reliable dataset for analyzing football-related language use in Serbian media.

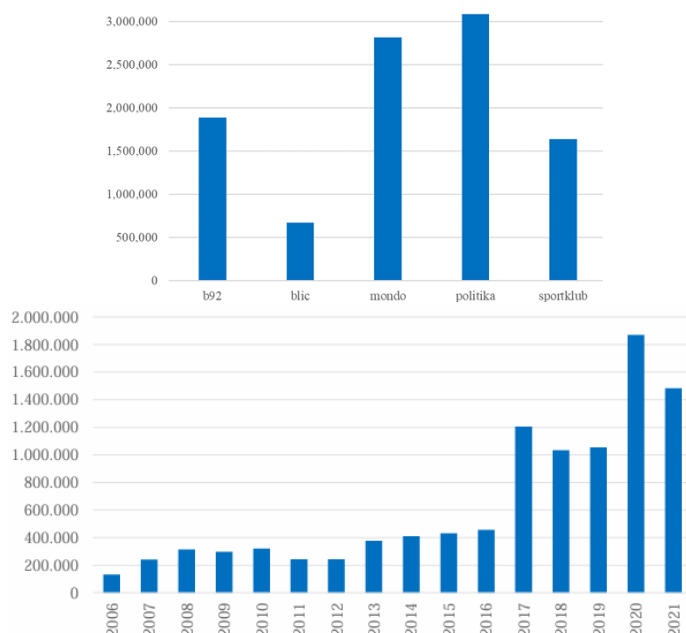


Figure 1. The partitions of Serbian football corpus *srFudKo* by web portals and years

Similarly, *esFudKo* corpus focuses on the football discourse in the Spanish language. It is compiled from articles published on the two major Spanish sports web-sites: Marca and Mundo Deportivo, tagged with TreeTagger Spanish tagset (<https://cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/spanish-tagset.txt>). The Spanish corpus contains 9,106,812 tokens, with 8,024,164 being words. The corpus *esFudKo* provides a comparable linguistic resource for analyzing Spanish football terminology, syntactic structures, and collocations (Lazarević & Kitanović 2024). By examining patterns within this dataset, we identify common linguistic features that define football reporting in Spanish-speaking media.

Together, these corpora serve as essential tools for comparative linguistic research. They enable cross-linguistic analysis of football

discourse, highlighting similarities and differences in the ways football is discussed, reported, and conceptualized in Serbian and Spanish. Additionally, these corpora facilitate automatic extraction of collocations, frequency analysis, and syntactic pattern recognition, contributing to a broader understanding of football-specific language in different linguistic and cultural contexts.

Both corpora used in this study have undergone data extraction methods and are available on two platforms: web-based access through Noske (<https://noske.jerteh.rs/>) (Kilgariff et al. 2014; Stanković et al. 2021), an online corpus analysis platform; and local access via TXM, a specialized corpus linguistics software for textual analysis (Heiden 2010;). Access to these corpora is restricted to authorized users, ensuring controlled use and data protection.

In this study, collocations are identified based on their mutual lexical-semantic attraction, determined through frequency measures and other statistical metrics applied within the corpora (Kitanović et al. 2021). This approach allows a systematic and empirical identification of word combinations that frequently occur together, ensuring that the extracted collocations are linguistically significant and relevant to the footballing language in both Serbian and Spanish.

4. Results

4.1. Analysis of Collocations of the Most Frequent Nouns in srFuDKo and esFuDKo

We now illustrate the analysis using the example: *utakmica*/match, a highly frequent noun in football discourse, analyzing its collocations helps us reveal typical word patterns and semantic relationships in both Serbian and Spanish football terminology. Collocations involving *utakmica* can be categorized into different structural patterns. For the pattern adjective + noun, illustrative examples are: *važna utakmica*/important match, *teška utakmica*/tough match, *prijateljska utakmica*/friendly match.

The pattern noun + noun can be illustrated with: *gol razlika*/goal difference, *revanš utakmica*/rematch. The pattern verb + noun can be illustrated with: *igrati utakmicu*/to play a match, *dobiti utakmicu*/to win a match, and *izgubiti utakmicu*/to lose a match. For the pattern noun + prepositional phrase, the illustrative examples are: *utakmica na stadionu*/match at the stadium, *utakmica pod reflektorima*/match under the floodlights, *utakmica za titulu*/match for the title.

The noun *utakmica*/match often appears in contexts describing competition, results, and game conditions. The collocations provide insight into how matches are characterized, whether by difficulty (*teška utakmica*/tough match), significance (*važna utakmica*/an important match), or type (*prijateljska utakmica*/friendly match). Additionally, the verb-noun combinations reflect actions within the game: winning (*dobiti utakmicu*/win the match), losing (*izgubiti utakmicu*/lose the match), or playing (*igrati utakmicu*/play the match). These structures are crucial for Computational linguistics applications such as machine translation, keyword extraction, and sentiment analysis in sports journalism.

Figure 2 presents the words (lemma) with strong connection calculated from articles as a list of collocates with the calculated statistical measures in *NoSketch engine* (Kilgarriff et al. 2014), maintained by the Language Resources and Technologies Society *JeRTeh*.

	Lemma	Cooccurrences	Candidates	T-score	MI
1	odigrati	1,392	6,662	36.78	6.14 ***
2	prvenstvo	1,368	9,808	36.20	5.56 ***
3	na	5,035	147,043	64.82	3.53 ***
4	prvi	1,738	32,008	39.41	4.20 ***
5	protiv	1,283	16,994	34.41	4.67 ***
6	tokom	816	4,267	28.12	6.01 ***
7	u	7,586	295,552	77.04	3.11 ***
8	ovaj	1,460	35,794	35.43	3.78 ***
9	za	2,652	89,281	46.36	3.33 ***
10	igrati	1,279	27,803	33.46	3.96 ***

	Lemma	Cooccurrences	Candidates	T-score	MI
11	kvalifikacioni	676	1,058	25.88	7.75 ***
12	poslednji	911	12,023	29.00	4.68 ***
13	posle	1,110	23,748	31.20	3.98 ***
14	.	8,949	403,033	81.97	2.91 ***
15	svaki	772	8,146	26.92	5.00 ***
16	nakon	756	7,589	26.68	5.07 ***
17	sa	2,016	69,548	40.31	3.29 ***
18	dva	996	20,306	29.65	4.05 ***
19	Srbija	1,025	25,046	29.70	3.79 ***
20	Portugal	601	3,470	24.10	5.87 ***

Figure 2. The Collocations panel in *srFudKo* on NoSketch engine for the word *utakmica*/match

In the Spanish language, similar collocations appear with the term *partido*/match. For adjective + noun pattern: *partido importante*/important match, *partido difícil*/tough match, *partido amistoso*/friendly match. For noun + noun pattern, we what can be seen is: *final del partido*/final match, *revancha del partido*/match rematch, while for Verb + Noun: *jugar un partido*/to play a match, *ganar un partido*/to win a match, and *perder un partido*/to lose a match. Despite language differences, the collocational patterns remain structurally similar, showing the universal nature of football terminology across languages.

4.2. The Adjective + Noun Pattern in srFuDKo and esFuDKo

The adjective + noun pattern is frequently used to describe properties or characteristics of an entity. This pattern, recognized as one of the most productive in a language, was analyzed in the *srFudKo* corpus by extracting occurrences using the CQL pattern: [srpos="ADJ"] [srpos="NOUN"]. The extraction process yielded 528 candidates for combinations with a frequency higher than 100, while 84 candidates had a frequency exceeding 500. In Table 1, the frequent adjective + noun collocations related to the footballing language are listed, illustrating typical word combinations within the language of football in Serbian.

Table 1. The frequent collocations for pattern adjective + noun in the corpus srFudKo

Bigram	Freq	Bigram	Freq	Bigram	Freq
Crvena zvezda	11050	prvi meč	1599	dobar rezultat	939
svetsko prvenstvo	4165	slobodni udarac	1478	drugi deo	834
drugo poluvreme	2810	prvi deo	1366	fudbalski klub	823
prvo poluvreme	2686	direktan plasman	1231	veliki klub	806
kazneni prostor	2317	grupna faza	1214	velika šansa	795
fudbalski savez	2254	istorijski trijumf	1121	dobar strelac	759
Evropsko prvenstvo	1808	prelazni rok	1090	sportski direktor	758
srpski fudbal	1807	mladi igrač	1051	dobra prilika	729
stručni štab	1708	bela tačka	1004	vezni red	727
žuti karton	1707	prvo mesto	990		
nacionalni tim	1647	crveni karton	963		

The most frequent nouns appearing in adjective + noun structure are: *utakmica*/match (67), *igrač*/player (66), *klub*/club (62), *tim*/team (54), *gol*/goal (50), *meč*/ match|game (49), *fudbaler*/footballer (42), *fudbal*/football|soccer (40), *godina*/year (34), *liga*/ league (31), *sezona*/season (31), *rezultat*/result (29), *šansa*/chance (27), *pobeda*/victory|win (26), *trener*/coach|manager (26), *ekipa*/team-squad (25), *igra*/game|play (24), *prilika*/ opportunity (23), *situacija*/situation (23), *napadač*/striker|attacker (22).

The most common adjectives modifying *utakmica*/match are: *prva*/first (978), *kvalifikaciona*/qualifying (663), *prijateljska*/friendly (514), *poslednja*/last|final (360), *važna*/ important (281), *teška*/tough|difficult (265), *velika*/big|major (252), *dobra*/good (241), *fudbalska*/football|soccer-related (219), *evropska*/European (189), *druga*/second (165), *naredna*/next|upcoming (155), *prvenstvena*/league-

related|championship (136), *odigrana/* played (132), *takmičarska/*competitive (117), *finalna/final*|decisive (114).

The analysis of the adjective + noun pattern in the *esFudKo* corpus used a different CQL pattern due to differences in part-of-speech tagging: [espos="ADJ"] [espos="NC|NMEA|NMON"]. This extraction yielded 271 candidates with a frequency higher than 50 and 724 candidates with a frequency exceeding 20. Table 2 presents the most frequent adjective + noun collocations in *esFudKo*.

Table 2. Most frequent collocations for pattern adjective + noun in corpus *esFudKo*

eslemma	F	eslemma	F	eslemma	F
Real Madrid	6798	bueno mejor versión	343	buen mejor equipo	228
Real Sociedad	1815	bueno partido	330	cristiano Ronaldo	228
próxima temporada	1063	mucha gente	328	gran grande partido	225
último partido	942	bueno mejor momento	318	próximo día	221
último año	879	nuevo entrenador	310	último encuentro	221
última jornada	781	última temporada	307	gran grande parte	214
último hora	538	buena sensación	300	buena parte	211
lateral derecho	478	última semana	297	alto nivel	210
bueno mejor jugador	456	buen resultado	278	ex-jugador	208
último día	436	buena mejor manera	275	media hora	206
buen jugador	435	último mes	259	gran jugador	205
último minuto	427	grande categoría	254	delantero central	203
recto final	426	mismo tiempo	254	último tiempo	200
buen momento	399	buen equipo	249		
grande máximo goleador	359	próxima semana	239		

In the *esFudKo* corpus, for the adjective + noun pattern where the noun is preceded by an adjective, the following examples stand out based on the number of collocates: *jugador/player* (14), *partido/match|game* (14), *club/club* (14), *equipo/team|squad* (14), *temporada/season* (12), *entrenador/coach|manager* (9), *momento/moment* (9), *gol/goal* (9), *futbolista/footballer|soccer player* (9), *año/year* (8), *ocasión/opportunity|occasion* (8), *jornada/matchday|round* (8).

The adjectives characterizing *jugador/player* are: *bueno|mejor/good|better* (456), *buen/good* (435), *ex/former* (208), *grande/great|big* (205), *nuevo/new* (177), *gran|grande/ great* (155), *propio/own* (91), *único/unique* (82), *mismo/same* (69), *joven/young* (63),

actual/ current|present (29), *último/*last|latest (29), *tanto/*so much (24), *excelente/*excellent (20).

When we compare the ten most frequent nouns for Serbian and Spanish that are part of those collocations, we can notice that they overlap. In Figure 3, with the same colour and background, translation equivalents are presented. The first Serbian noun is *utakmica*/match (67), and the Spanish *jugador*/player (14) is in first place. More precisely, the first four have the same frequency in the Spanish dataset.

Rang	Srpski	Rang	Španski
1	<i>utakmica</i> (67)	1-4	<i>jugador</i> (14)
2	<i>igrač</i> (66)	1-4	<i>partido</i> (14)
3	<i>klub</i> (62)	1-4	<i>club</i> (14)
4	<i>tim</i> (54)	1-4	<i>equipo</i> (14)
5	<i>gol</i> (50)	5	<i>temporada</i> (12)
6	<i>meč</i> (49)	6	<i>entrenador</i> (9)
7	<i>fudbaler</i> (42)	7	<i>momento</i> (9)
8	<i>fudbal</i> (40)	8	<i>gol</i> (9)
9	<i>godina</i> (34)	8	<i>futbolista</i> (9)
10	<i>liga</i> (31)	10	<i>año</i> (8)

Figure 3. The most frequent nouns in the adjective + noun pattern collocation *srFuDKo* and *esFudKo*

4.3. Other Patterns with Nouns within Corpora *srFudKo* and *esFudKo*

The analysis of the Serbian corpus *srFudKo* and the Spanish corpus *esFudKo* reveals interesting differences in the use of the words *lopta*/ball and Spanish *balón*/ball, which manifest through distinct collocation types and distribution patterns. In both corpora, *lopta* and *balón* most frequently collocate with nouns related to different elements of the game. However, there are significant differences in these nouns and their functional roles.

In *srFudKo*, *lopta*/ball is most commonly associated with nouns that describe physical aspects of the game and players, such as *mreža*/net, *glava*/head, *korner*/corner, *noga*/leg, *šesnaesterac*/penalty area, *golman*/goalkeeper, *udarac*/shot, *teren*/field, and *igrač*/player. These collocations clearly describe direct aspects of football and the physical actions involved, emphasizing players' interactions with the ball.

In contrast, within *esFudKo*, *balón*/ball frequently appears in collocations related to broader game concepts and ball possession, such as *posesión*/possession, *control*/control, *dominio*/dominance, *juego*/game|play, and *acción*/action. While there are also collocates like *pie*/foot, *área*/penalty area, and *gol*/goal, with greater focus on tactical aspects of the game suggests that the Spanish language of football emphasizes conceptual and strategic elements more than physical actions.

In the research on football-related collocations, the following queries focus on the noun *lopta*/ball, as an example, identifying its syntactic and semantic relationships with surrounding words. The first two queries, [srlemma = "lopta" & srpos = "NOUN"] [srpos != "PUNCT"] and [srpos != "PUNCT"] [srlemma = "lopta" & srpos = "NOUN"], retrieve instances where *lopta* appears before or after any word that is not punctuation, ensuring a clean extraction of collocational structures. The third query, [srlemma = "lopta" & srpos = "NOUN"] [srpos = "ADJ"]* [srpos = "NOUN"]+, captures multi-word noun phrases involving *lopta*, such as *lopta za igru*/game ball or *lopta u mreži*/ball in the net, reflecting common football terminology. The final query, [srpos = "VERB"]+ []{0,2} [srlemma = "lopta" & srpos = "NOUN"], focuses on verb + noun collocations, allowing for the identification of structures like *šutirati loptu*/kick the ball, *uhvatiti loptu*/catch the ball, or *dodati loptu*/pass the ball. Figure 4 presents some results extracted from these queries that help uncover syntactic patterns and lexical preferences within Serbian football discourse, providing insights into phraseology, word order, and semantic roles in sports language.

In Spanish, the noun *balón*/ball most frequently appears with verbs such as *tocar*/touch, *parar*/stop, *robar*/steal, *mandar*/send, *perder*/lose, *controlar*/control, *rodar*/roll, and *despejar*/clear. While these verbs also refer to physical actions, there is a stronger emphasis on ball control (*controlar*, *parar*, *tocar*), confirming that the Spanish football language places greater importance on possession and tactical play compared to the Serbian football language.

A particularly interesting contrast emerges when examining the most common collocations in each language. In Serbian, *lopta*/ball is most frequently associated with the verb *poslati*/send, emphasizing a direct, action-based approach. In Spanish, however, *balón*/ball is most frequently linked with the adjective *oro*/gold, forming the syntagm *Balón de Oro*, referring to the most prestigious award for the world's best football player. It has no direct equivalent in Serbian in terms of awards, while its title is used through a literal translation – *Zlatna lopta*.

[srlemma = "lopta" & srpos = "NOUN"] [srpos != "PUNCT"]		[srpos != "PUNCT"] [srlemma = "lopta" & srpos = "NOUN"]		[srlemma = "lopta" & srpos = "NOUN"] [srpos = "ADJ"] [srpos = "NOUN"]+		[srpos = "VERB"]+ [] [0,2] [srlemma = "lopta" & srpos = "NOUN"]	
srlemma	F	srlemma	F	word	F	word	F
lopta u	2823	jesam lopta	1779	loptu glavom	205	poslao loptu	713
lopta jesam	1902	poslati lopta	843	loptu pravo	37	ubacio loptu	248
lopta na	1209	zahvatiti lopta	349	loptu rukom	28	zahvatio loptu	247
lopta i	1097	sa lopta	330	loptu grudima	23	primio loptu	189
lopta do	352	ubaciti lopta	311	loptu golmanu	19	zakucao loptu	170
lopta sa	309	i lopta	309	loptu rivalu	15	dobio loptu	159
lopta koji	303	na lopta	292	loptu protivniku	14	šalje loptu	155
lopta iz	282	izbaciti lopta	291	loptu nogom	13	izbacio loptu	141
lopta za	272	dug lopta	282	loptu petom	10	dodao loptu	139
lopta se	269	a lopta	263	loptu gostima	9	izgubio loptu	125
lopta otići	215	do lopta	254	lopte glavom	8	izbio loptu	118
lopta glava	213	posed lopta	245	loptu okrenut leđima	8	vratio loptu	115
lopta ka	209	primiti lopta	238	loptu saigračima	8	prosledio loptu	98
lopta posle	194	zlatan lopta	231	lopta celim obimom	7	smestio loptu	98
lopta poslati	161	se lopta	220	loptu levom nogom	7	odbio loptu	90
lopta od	154	izbiti lopta	213	loptu '	7	oduzeo loptu	90
lopta pored	132	ali lopta	204	loptu igračima	6	spustio loptu	89
lopta preko	127	odbijen lopta	204	loptu tik	6	prihvatio loptu	87
lopta ici	122	zakucati lopta	203	loptu pravo	5	uzeo loptu	77
lopta pogoditi	120	dodati lopta	184	loptu '	5	zakačio loptu	73
lopta iza	116	izgubiti lopta	182	loptu igrači	5	promašio loptu	71
lopta po	115	slati lopta	181	lopte '	4	plasirao loptu	63
lopta završiti	114	s lopta	181	loptu defanzivcima	4	sproveo loptu	63
lopta kroz	111	dobiti lopta	175	loptu igraču	4	prebacio loptu	60
						stignuti do lopta	83

Figure 4. The examples of multiword terms retrieved from *srFudKo*

4.4. Beyond Frequency: Context and Semantic Nuances in Football Collocations

Sports fans incorporate sports-related expressions into everyday speech, gradually introducing them to the wider public. These phrases often preserve their original meaning as they transition from the football field, through journalists and fans, to common usage (Lavrić 2008). In addition to analyzing the frequency of occurrence, this study examines other factors such as semantic proximity and context of use. Some collocations in football terminology carry specific meanings or appear only in particular situations, reflecting tactical nuances and player strategies.

This is shown in the following examples:

1) *dati gol iz mrtvog ugla*/score a goal from a dead angle represents that this phrase is used in situations where a player scores from an unexpected or difficult position, often from an angle where it seems nearly impossible for the ball to enter the goal;

2) *izvući penal*/draw a penalty– while the referee can award a penalty kick. The phrase *izvući penal* implies that the player skillfully exploited a situation or initiated contact with an opponent in a way that

convinced the referee to award a penalty (this highlights the tactical and psychological aspects of the game);

3) *igrati bunker*/play bunker defense – a defensive strategy where a team withdraws into a compact defensive formation, avoiding offensive play and waiting for a counter-attack opportunity (the primary goal is to protect their own net, often used against stronger opponents or when defending a lead). These collocations demonstrate how football language extends beyond the literal meanings, incorporating strategic, physical, and tactical elements, making them essential for understanding football discourse in Serbian.

We also examine whether previously identified collocations are understandable to the general public, particularly to those who do not watch football and are unfamiliar with football-specific language. If a speaker from the general public can understand and use these phrases outside a sports context, it indicates that the collocations have transcended their original football domain, becoming part of the public domain. For example:

1) *igrati na sigurno*/play it safe in football. This phrase refers to a strategy of avoiding risks, such as a team focusing on defense rather than aggressive attacks, while in everyday language, it describes a cautious approach to decision-making: "I'd rather play it safe and invest in low-risk funds.";

2) *autogol*/own goal. This term is originally a football term describing a situation where a player accidentally scores against their team. It is now widely used to describe self-inflicted mistakes in any context: "...by rejecting that job offer, he scored an own goal".

3) *igrati na domaćem terenu*/to play on home ground (at home). In football, this denotes a match played in the team's own stadium or home venue, where the team typically has an advantage due to familiarity with the pitch, support from local fans, and absence of travel fatigue. In the public domain, it means that someone is in a familiar, favorable environment, where they have an advantage: "... The president insisted the meeting be held in Belgrade because he wants to play on home ground." with the meaning: using a favorable location to gain strategic advantage in negotiations.

These examples illustrate how deeply ingrained football terminology has become within the General speech, showcasing the way in which the language of sport, especially football, influences the broader cultural and linguistic expression.

5. Conclusion

The language of sport in both Serbian and Spanish is predominantly shaped by football terminology, which has transcended its original domain, becoming an integral part of the General public discourse. Certain football-related terms are equally prevalent in both languages, reflecting the global reach and cultural significance of football. The analysis of the noun *utakmica*/match and its collocations highlight the lexical-semantic connections within the language of football. Understanding these patterns enhances linguistic research, sports journalism, and NLP applications related to sports language. Comparing Serbian and Spanish football terminology further deepens insights into how football-related words function across languages. This study contributes to these insights by analyzing the connections between collocations and multi-word football terms, which is particularly important when multi-word terms contain collocates with a clear meaning within the footballing domain. This research showcases the frequency of collocations and their possible combinations, highlighting recurring patterns that define football discourse in both languages.

Acknowledgment

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

References

- [1] Heiden, S. (2010). The TXM platform: Building open-source textual analysis software compatible with the TEI encoding scheme. 24th Pacific Asia conference on language, information and computation, 2(3), 389–398.
- [2] Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography*, 1(1), 7–36.
- [3] Krstev, C., Stanković, R., Obradović, I., & Lazic, B. (2015). Terminology Acquisition and Description Using Lexical Resources and Local Grammars. *TIA*, 81–89.
- [4] Kitanović, O., Stanković, R., Tomašević, A., Škorić, M., Babić, I., & Kolonja, L. (2021). A data-driven approach for raw material terminology. *Applied Sciences*, 11(7), 2892.
- [5] Lavric, E. (2008). The linguistics of football (Sv. 38). BoD–Books on Demand.

- [6] Lazarević, J. V. (2024). Tekstometrija u korpusnoj linivistici: srpski fudbalski korpus *srFudKo*. Nasleđe Kragujevac, XXI(59), 111–124.
<https://doi.org/10.46793/NasKg2459.111L>
- [7] Lazarević, J., & Kitanović, O. (2024). Building the Dictionary of Football Terminology Through Data-Driven and Ontolex Model. *Infotheca - Journal for Digital Humanities*, 24(1), pp. 29–52, DOI 10.18485/infotheca.2024.24.1.2.
- [8] Lazarević, J., Stanković, R., Škorić, M., & Rujević, B. (2023). Football terminology: Compilation and transformation into OntoLex-Lemon resource. U S. Carvalho, A. F. Khan, A. O. Anić, B. Spahiu, J. Gracia, J. P. McCrae, D. Gromann, B. Heinisch, & A. Salgado (Ur.), *Proceedings of the 4th Conference on Language, Data and Knowledge* (str. 634–645). NOVA CLUNL, Portugal.
<https://aclanthology.org/2023.ldk-1.69>
- [9] Levin, M. (2008). Hitting the back of the net just before the final whistle: High-frequency phrases in football reporting. *The linguistics of football*, 143–155.
- [10] Pejović, A. (2007). Frazeologija u jeziku španskih novina. *Nasleđe*, 4(8), 155-167.
- [11] Pejović, A. (2015). Kontrastivna frazeologija španskog i srpskog jezika. *Filološko-umetnički fakultet*.
- [12] Prčić, T. (2016). *Semantika i pragmatika reči*. Filozofski fakultet.
- [13] Stanković, R., Sandrih, B., Krstev, C., Utvić, M., & Škorić, M. (2020). Machine learning and deep neural network-based lemmatization and morphosyntactic tagging for Serbian. *LREC 2020-12th International Conference on Language Resources and Evaluation, Conference Proceedings*, 3954–3962.
- [14] Stanković, R., Krstev, C., Todorović, B. Š., & Škorić, M. (2021). Annotation of the Serbian ELTeC Collection. *Infotheca–Journal for Digital Humanities*, 21(2), 43-59.

Kontrastivna analiza sintaksičkih obrazaca u komparabilnim korpusima fudbala na španskom i srpskom jeziku

Jelena Lazarević, Olivera Kitanović

Sažetak

Ovaj rad istražuje pojam kolokabilnosti kao načina na koji se leksičke jedinice kombinuju sa različitim kategorijama reči kako bi formirale veće jedinice, korišćenjem analize semantičkih i sintaksičkih principa srpskih i španskih fudbalskih termina korišćenjem uporedivih korpusa: *srFudKo* i *esFudKo*, razvijenih u okviru doktorske disertacije Jelene Lazarević: *Jezičke odlike diskursa novih medija o fudbalu: kontrastivna analiza na srpskom i španskom jeziku*. Korpus *srFudKo* obuhvata 10.100.553 tokena sa pet srpskih veb-sajtova sa vestima iz fudbala, dok se *esFudKo* sastoji od 9.106.812 tokena sa dva španska fudbalska veb-sajta. Oba korpusa se analiziraju korišćenjem metoda korpusne lingvistike korišćenjem CQL (Corpus Query Language) i tekstometrijske analize. Studija ispituje kolokacije na osnovu frekvencije i semantičke privlačnosti, identifikujući različite tipove, kao što su: pridev + imenica, imenica + imenica, glagol + imenica, prilog + pridev, glagol + predložka fraza i glagol + prilog. Proces ekstrakcije kolokacije oslanja se na sintaksičke obrasce i analizu frekvencija. Osim frekvencije, istraživanje razmatra semantičku bliskost, kao i kontekst, procenjujući da li kolokacije nose specifična značenja i da li su razumljive široj javnosti. Kada se široko shvate, kolokacije iz posmatranih korpusa prevazilaze fudbalski domen. Ovo istraživanje takođe ispituje veze između kolokacija i višečlanih termina, pokazujući kako termini specifični za fudbal doprinose širem razumevanju sportske jezičke strukture u srpskom i španskom jeziku.

Ključne reči: fudbal, korpusi, terminologija, kolokacije, srpski, španski

Синтаксичка и семантичка анализа глагола *оштетити* и *уништити* из угла теорије семантике оквира

Научни рад

DOI: 10.18485/judig.2025.1.ch11

Марина Баги¹  0009-0006-9410-7862

Апстракт

У раду се представља анализа глагола *оштетити* и *уништити* из угла теорије семантике оквира. Полазећи од претпоставке да синтаксички и семантички феномени међусобно корелирају, за теоријску основу рада бира се когнитивносемантички приступ Ч. Филмора који експлицитно повезује значења неке речи са синтаксичким контекстима у којима се она јавља. Основна теза теорије семантике оквира јесте да значења неке сублексеме треба посматрати у вези са семантичким оквирима – схематским приказима појмовних структура и образаца веровања који леже у основи значења речи. Ова теорија у основи је Фрејмнета, електронске лексикографске базе података која је заснована на корпусу, а у оквиру које су приказани оквири које различите лексеме призивају.

Истраживање је засновано на шест стотина примера употребе лексема *оштетити* и *уништити* ексцерпираних из електронског *Корпуса савременог српског језика (СрпКор2021)* са циљем да се опише и испита њихово синтаксичко и семантичко понашање, те да се прикаже њихова употреба. Стога се описују оквири које призивају поменути глаголи, као и елементи оквира, а изводе се и закључци у вези са синтаксичко-семантичким интерфејсом. Уз то, будући да су у питању глаголи сличне семантике, циљ је да се истраже сличности и разлике између њих, што се чини употребом алата на платформама *Sketch Engine* и *No Sketch Engine*.

¹ Институт за српски језик САНУ, marina.bagi@isj.sanu.ac.rs

Кључне речи: теорија семантике оквира, синтакса, семантика, глагол *оштетити*, глагол *уништити*, обрада природних језика, српски језик.

1. Увод

Теорија семантике оквира представља когнитивносемантички приступ анализи лексичког значења који је развио Чарлс Филмор заједно са својим сарадницима. Централна идеја овог приступа јесте да значења речи треба описивати у вези са семантичким оквирима – схематским приказима појмовних структура и образаца веровања који леже у основи значења речи (Fillmore et al. 2003, 235). Термином оквир² означен је својеврстан систем појмова који су повезани на такав начин да је за разумевање једног појма неопходно разумети читаву структуру чији је он део (Fillmore 1982, 111), а једна од основних теоријских поставки семантике оквира јесте да оквире призивају³ речи одређених врста (у првом реду глаголи, а потом именице, придеви и прилози).

Поменути приступ у фокус ставља употребу речи. У вези са тим, Ч. Филмор (1969, 94) писао је да су проучаваоци, постављајући питање о значењу неког облика, заобилазили кључну ствар – питање *употребе* речи, односно шта је оно што морамо да знамо како бисмо одређени облик адекватно користили и како бисмо разумели друге људе када га употребљавају (курзив наш).

Особит значај теорије семантике оквира огледа се у томе што представља теоријско-методолошки оквир који пружа могућност да се експлицитно повежу значења речи са синтаксичким контекстима у којима се те речи појављују (Atkins et al. 2003a, 254). У литератури се све више истиче значај повезивања традиционалне синтаксе и лексикографије, а интеграција синтаксичког система и лексикона сматра се важним кораком у модернизацији граматичког и лексикографског описа српског стандардног језика (Поповић 2003, 201, 204). Интегрални приступ језику подразумевао би усмеравање синтаксичких истраживања и на сублексеме, односно конкретне лексеме у одређеном значењу (Поповић 1997, 208; Стипчевић 2010, 175), док би требало да приказ лексикона буде интегралан, те да садржи све синтаксички релевантне информације (Поповић 1997, 481; Поповић 2003, 204; Стипчевић 2010, 175).

² Оквир функционише као општи назив за појмове *схема*, *сценарио*, *когнитивни модел*, *народна теорија* и сл. (Fillmore 1982, 111).

³ Глагол *призвати* употребљава се као еквивалент енглеског глагола *to evoke* у преводима радова Ч. Филмора (Rasulić i Klikovac 2014, 79), као и у досадашњој литератури о могућностима примене теорије семантике оквира (Марковић и др. 2021, 9), што је решење за које се опредељујемо и у овом раду.

Идентификовање и дефинисање значења полисемичних речи представља значајан изазов за лексикографе и лингвисте. Будући да се међусобно прожимају, значења се често не могу јасно разграничити, што отежава њихово прецизно издвајање. Ипак, примена одговарајућих теоријских приступа и формалних метода може пружити јасније смернице у овом процесу, на пример, у одређивању оптималне листе значења која ће обухватити потврђене примере из корпуса (Fillmore and Atkins 2000, 100). Теорија семантике оквира додатно омогућава да се лексикографски опис обогати корпусним подацима који би иначе могли остати незапажени, да се препознају карактеристике кључне за опис значења речи и да се издвоје у великој количини података доступних у савременим корпусима. Према речима С. Еткинс, без довољно јасне и научно утемељене формалне методе, корпусна анализа би остала импресионистичка (Atkins 1994, 45).

Семантика оквира своју практичну примену пре свега налази у Фрејмнету (енгл. *FrameNet*), електронској лексикографској бази података⁴ која је заснована на анотацији примера неке кључне речи у аутентичном тексту. Подаци се црпу из корпуса комбинацијом ручних и аутоматских процедура, а те информације о повезаним семантичким и синтаксичким особинама речи представљају се кроз различите извештаје⁵ који су доступни на интернету (Fillmore et al. 2003, 235).

Централни елемент лексичке анализе у семантици оквира и Фрејмнету представља опис *лексичке јединице*, под којом се подразумева лексема у једном од својих значења (Ruppenhofer et al. 2016, 7). Уместо термина *лексичка јединица* користимо термин *сублексема*, као што је уобичајено у србистици.

1.1. Претходна истраживања

Лексеме *оштетити* и *уништити* нису често биле предмет разматрања у српској лингвистици. Глагол *уништити* проучавао је В. Миљковић (2014), који је анализирао његову семантику из угла теорије менталних простора, која глаголе такође посматра као својеврсне оквири, односно менталне просторе који се састоје од семантичких улога и релација (143). Б. Хлебец (Hlebec 2023), са друге

⁴ Рад на пројекту Фрејмнета започет је 1997. године на Међународном институту за рачунарске науке у Берклију, а база је доступна на следећем линку: <https://berkeleyfn.framenetbr.ufjf.br>.

⁵ Под поменутих извештајима углавном се подразумевају синтаксички и семантички описи валенце, добијени вишеслојном анотацијом примера из корпуса. Представљање ових резултата превазилази оквири нашег рада.

стране, компаративно посматра енглеске⁶ и српске глаголе оштећења (међу њима и глагол *оштетити*), које је дефинисао колокацијском методом. Закључке ових радова износићемо у наредним одељцима.

Колико нам је познато, до сада није било много речи о примени теорије семантике оквира на материјалу српског језика. Изузетак представља рад Марковић и др. (2021), који приказује резултате прелиминарних истраживања у вези с могућностима примене ове теорије и пружа детаљан приказ фаза и процеса у раду на Фрејмнету, на примеру неколико оквира из домена ризика.

1.2. Методологија истраживања и корпус

Истраживање је започето претрагом глагола *оштетити* и *уништити* на платформи *Sketch Engine*, на којој су најпре коришћени алати *Word Sketch* и *Word Sketch Difference* како би се добила општа слика о употреби ових лексема (т. 3). На основу добијених резултата, који су потом проверени претрагама у CQL-у на већем *Корпусу савременог српског језика* (*СрпКор 2021*; <https://noske.jerteh.rs/>), ексцерпирали смо 600 реченица у којима су употребљене лексеме *оштетити* и *уништити*. Највећи део потиче са корпуса *СрпКор 2021*, а мањи број добијен је претрагом на *Гуглу* и на корпусу *ПДРС 1.0*.

Све примере смо анализирали пратећи приручник за израду Фрејмнета (Ruppenhofer et al. 2016). Најпре смо ексцерпирани материјал разврстали према томе који оквир призива сублексема употребљена у сваком од примера, а онда смо настојали да призване оквири опишемо. Да би се то учинило, неопходно је да се формулише ситуација означена оквиром и наведу улоге учесника у датој ситуацији (Atkins et al. 2003a: 252). Улогама учесника одговарају централни и периферни елементи оквира. Централним елементима сматрају се они који врше неку од базичних граматичких функција, чија је синтаксичка реализација обавезна, а без којих оквир као такав не би могао да постоји. Са друге стране, периферни су они елементи који синтаксички, грубо речено, одговарају одредбама (нпр. место, време, начин, средство...).⁷

⁶ Уп. и радове који се ослањају на анализу корпуса и колокација приликом проучавања глагола 'damage' и 'destroy' (Song 2021), односно блиских синонима глагола 'destroy' (Ivanova and Medvedeva 2023).

⁷ Семантика оквира разликује и екстратематске и централно неизражене елементе оквира (Ruppenhofer et al. 2016, 24–25). Њима се у овом раду нећемо бавити.

Током анализе смо све оквири допуњавали одговарајућим информацијама.⁸ Смернице приликом описа оквира пружиле су, поред генерисаних скица речи, и дефиниције ових лексема у *Речнику српскога језика*, *Речнику српскохрватскога књижевног језика* и *Речнику српскохрватског књижевног и народног језика САНУ*.

Један од основних циљева рада је да се, кроз описе оквира, прикаже полисемична структура лексема *оштетити* и *уништити*, али и да се ослика њихово синтаксичко понашање, да се, затим, уоче сличности и разлике међу овим глаголима, те и да се представи типично оштећивање и типично уништавање, као и прелазни случајеви који се међу њима јављају.

2. Анализа

На основу података из корпуса установили смо да се лексема *оштетити* повезује са оквирима Оштећивање⁹, Изазвати_емоцију и Ускраћивање, а лексема *уништити* са оквирима Уништавање, Убијање, Победити_противника, Стимулисати_емоцију, Изазвати_емоцију и Довести_у_тешко_стање.¹⁰ Представимо укратко сваки од њих тако што ћемо у деловима приказати оквири онако како би изгледали у Фрејмнету¹¹.

2.1. Оштећивање

Ситуација оштећивања подразумева постојање двају централних елемената оквира: АГЕНСА (или УЗРОКА) и ПАЦИЈЕНСА. АГЕНС или УЗРОК утичу на ПАЦИЈЕНС тако да он (или његов део) заврши у неканонском, непожељном стању¹². Овај оквир повезан је са другим, сличним оквирима (Уништавање, Изазвати_распарчавање и Покварити). Интересантно је да су радња и резултат у вези са глаголом *оштетити*

⁸ Процес рада на дефинисању оквира и његових елемената није линеаран. Наиме, сем што нова фаза анализе може да донесе другачија сазнања, тако и анализа других лексема које призивају тај оквир може да изискује измену у његовом опису.

⁹ Прегледности ради, називе оквира ћемо писати фонтом *Courier New*, док ћемо називе елемената оквира наводити малим верзалом.

¹⁰ Фрејмнет база садржи и информације о различитим односима (Ruppenhofer et al. 2016, 79–85) који постоје међу оквирима (попут *Наслеђивања*, *Коришћења* и *Угла_гледања*). Њима се у овом раду нећемо моћи посветити.

¹¹ Када год је то било могуће, током анализе смо пратили одговарајуће оквири који су доступни на Фрејмнету, али смо их прилагодили стању у српском језику. Детаљније поређење еквивалентних српских и енглеских глагола није нам било у фокусу, будући да таква анализа превазилази оквири нашег рада.

¹² До сличног закључка долази Хлебес, који сматра да глаголи оштећења деле дефиницију 'изазвати да буде у лошем стању' (Hlebec 2003, 1).

много мање спецификовани у односу на неке друге глаголе, као што су *огребати*, *стругати*, *улубити* и сл.

После дефиниције следе три аотирана примера¹³. АГЕНС је у њима свесни ентитет који врши намерну радњу, док УЗРОК најчешће представља догађај или појам који заузима место у времену.¹⁴ У улози ПАЦИЈЕНСА овде су енглески циц, бродови и централни нервни систем.

Оштећивање

Дефиниција:

Агенс или **Узрок** утичу на **Пацијенс** тако да **Пацијенс** (или неки део на њему) заврши у неканонском стању. Често је ово неканонско стање непожељно, а неке лексичке јединице (означене семантичким типом Негативно) посебно указују на то да је **Пацијенс** негативно погођен. Неке лексичке јединице овог оквира указују на који начин је **Пацијенс** измењен (нпр. *стругати*) и у неким случајевима дају детаљније назнаке утицаја на **Пацијенс** (нпр. *засећи*), а у другим случајевима може бити наведено и једно и друго (*поцепати*). За други скуп лексичких јединица радња и њен резултат много су мање спецификовани (нпр. *оштетити*). Упоредити овај оквир са оквирима Уништавање, Изазвати, распарчавање и Покварити.

Непрестано одгуркујући нестрпљивог Воробјањина, **Остап** пажљиво оголи столицу, да **не би оштетио** енглески циц са цветићима.

Поплављен је део старог средишта града са старим каменим кућама, море се излило на риву и саобраћајнице, а **јак ветар** **оштетио** је и **неколико бродова** **наносећи** их на обалу.

Ендосулфан може да **оштети** **централни нервни систем** па чак и да изазове смрт, објаснили су филипински лекари.

Слика 1: Оштећивање (1).

¹³ Сваком елементу оквира додељује се одређена боја, а црном бојом означава се анализирана сублексема. После дефиниције елемента оквира наводи се по један аотиран пример за сваки елемент.

¹⁴ Елементи оквира се дефинишу на начин који је специфичан за сваки оквир (енгл. *frame-specific*), премда се могу посматрати као примери општијих семантичких улога (Марковић и др. 2021, 11).

Централни елементи:

Агенс: Свесни ентитет, обично особа, који врши намерну радњу која доводи до тога да **Пацијенс** трпи одређену штету.

Семантички тип: разуман, свестан

Искључује се: Узрок

У наредна три века **Турци** **су** два пута тешко **оштетили** манастир.

Узрок: Догађај или други ентитет који доводи до оштећења **Пацијенса**.

Искључује се: Агенс

НАТО бомбардовање 1999. године **је** директно **оштетило** водоводе у многим градовима, као и хемијска постројења, што је проузроковало загађење воде и тла.

Пацијенс: Ентитет који бива оштећен због активности **Агенса** односно због **Узрока**.

Тако су почетком септембра у српском селу Клокот на путу Урошевац – Гњилане експлодирале две mine које **су оштетиле** **две куће** и уништиле шталу.

Слика 2: Оштећивање (2).

Осим човека, у улози АГЕНСА се може наћи појам у метонимијској релацији са човеком (*НАТО алијанса*), а УЗРОК још представљају и природне појаве/катастрофе (*пљусак, поплава, град; ватра*), предмети којима управља човек (*аутобус, ручна бомба, метак*), као и неке врсте материја или једињења (*киселина, алкохол*).

ПАЦИЈЕНС врло често представљају локације, грађевине и већи физички објекти (*кућа, зграда, манастир, аутомобил*), затим медицински и здравствени ентитети (*мрежњача, кожа, зуби, вилица, вид*), биљке (*род јечма, купус, пшеница*), предмети (*тапете, циц, брава*) и информације (*гласачки листић*).

Периферни елементи оквира такође су важни чиниоци за разумевање целокупне слике коју може да подразумева ситуација оштећивања. У анализираном корпусу издвојили су се следећи елементи (сл. 3–4).

МАЛЕФИЦИЈАР је ентитет ком припада ПАЦИЈЕНС и који посредно трпи штету (*јој*).

КАРАКТЕР_КРАЈЊЕГ_СТАЊА је опис стања ПАЦИЈЕНСА након оштећивања, често реализован у виду прилошке јединице. СТЕПЕН до ког је ПАЦИЈЕНС оштећен означен је последичном зависном клаузом (*толико да је било немогуће поправити мотор на терену*).

САДРЖАВАЈУЋИ_ДОГАЂАЈ је елемент оквира којим су означени догађај или стање ствари који обухватају и ситуацију оштећивања. Најчешће је изражен именичком јединицом у ППК у + локатив, као у примеру *У бомбардовању је уништено и оштећено 25.000 стамбених објеката*, у ком се догађај или активност бомбардовања

концептуализује као садржатељ (Klikovac 2000, 53). Овај елемент јавља се онда када је сублексема *оштетити* употребљена у пасивној форми, те нема израженог АГЕНСА. Ако се догоди да неки централни елемент оквира није језички спецификован, као у овом случају, он ће ипак бити забележен у бази будући да постоји у појмовној структури. Такав случај назива се нултим појављивањем¹⁵ (енгл. *null instantiation*).

Периферни елементи:

Малефицијар: Ентитет који, такође, посредно трпи штету.

Међутим, топло време које влада ових неколико дана није нормално за ово доба године и неће бити добро ако пшеница сада нагло озелени, јер ће касније, док се буде развијала, опет захладити, и може се десити да **то** и мањи мразеви **оштете** лише.

Карактер крајњег стања: Опис стања **Пацијенса** након оштећења, укључујући опис тежине и трајности оштећења.

Током борбе Стојадиновићу **је** **припремено** **оштећен** слух пошто је нападач испалио два хица.

Степен: Степен до којег је **Пацијене** оштећен.
Семантички тип: степен

Београђанин Божо Ристић завршио је такмичење на Дакар релију, пошто је у седмој етапи воженеј **од Мендосе у Аргентини до Валпариса у Чилеу** пао са мотора и **оштетио** га **толико да је било немогуће поправити мотор на терену**.

Садржавајући догађај: Овај елемент оквира означава догађај који траје или стање ствари које постоји у времену и обухвата период у коме се одвијају догађаји или стања ствари изражени циљном речју.

У бомбардовању **је** уништено и **оштећено** 25.000 стамбених објеката, онеспособљено 470км путева, 595км пруга, оштећено је 14 аеродрома, 19 болница, 20 домова здравља, 18 вртића, 69 школа, 176 споменика културе и 44 моста, док је 38 разорено. **CNI**

Околност: Овим елементом оквира се назначавају околности у оквиру којих се врши оштећивање.

АМСС возачима скреће пажњу и на тзв. ударне рупе које, **при нешто већим брзинама**, могу да **оштете** возило, али и да угрозе безбедност возача и његових сапутника.

Место: Место на ком се одвија активност у којој **Пацијене** који трпи оштећивање.
Семантички тип: лоцираност_релативна

У области северно од великог туристичког центра Орланде, у олуји **је оштећено** или уништено више од 500 кућа, зграда и цркава.

Слика 3: Оштећивање (3).

¹⁵ У нашим примерима забележен је случај CNI (енгл. *Constructionally Null Instantiation*), који представља нулто појављивање одређено граматичком конструкцијом (нпр. изостављен агенс пасивних клауза), али и случај анафорског изостављања – DNI (енгл. *Definite Null Instantiation*). Семантика оквира разликује и INI – индефинитно нулто појављивање (енгл. *Indefinite Null Instantiation*) (Atkins et al. 2003a, 351–354).

ОКОЛНОСТ назначава прилике у оквиру којих се врши оштећивање (*при нешто већим брзинама*), а може бити наглашено и зашто АГЕНС односно УЗРОК оштећује ПАЦИЈЕНС. Тај елемент се назива ОБЈАШЊЕЊЕ и често је изражен именичком јединицом у ППК из + генитив (*из идеолошких разлога*). ИНСТРУМЕНТ је физички ентитет који АГЕНС користи како би нанео штету ПАЦИЈЕНСУ (*са два пројектила*), а СРЕДСТВОМ се означава намерна радња коју АГЕНС врши, а којом се остварује оштећивање (*употребом разнородних дрога*).

У слици ситуације може бити присутан и НАЧИН (*највероватније*), а поред МЕСТА и ВРЕМЕНА, који се не разликују много међу оквирима, може се појавити и ФРЕКВЕНЦИЈА, којом се назначује колико се често врши оштећивање (*два пута*).

Објашњење: Стање ствари на које Агенс реагује у процесу оштећивања.
Семантички тип: стање_ствари

Нажалост, после рата камен темељац је оштећен из идеолошких разлога, чак је украдена и спаљена Повеља која се налазила испод њега.

Инструмент: Ентитет којим управља Агенс, који ступа у интеракцију са Пацијенсом да би извршио оштећивање.
Семантички тип: физички_ентитет

Терористи су са два пројектила из ручног бадача оштетили и цркву Свети Никола у Тетову.

Средство: Намерна радња коју је извршио Агенс, а којом се остварује оштећивање.
Семантички тип: стање_ствари

Дејвид Боуви је морао да чита стихове својих песама током концерта јер је употребом разнородних дрога дуго низ година знатно оштетио своју меморију.

Начин: Сваки опис радње оштећивања која није обухваћена специфичнијим ЕО, укључујући епистемичке модификације (*вероватно*), силу (*тешко*, *слабо*), секундарне ефекте (*тихо*, *гласно*) и опште описе који упоређују догађаје (*на исти начин*). Може указивати на истакнуте карактеристике Агенса које такође утичу на радњу (*дрско*, *хладно*, *намерно*, *жељно*, *пажљиво*).
Семантички тип: начин

Према нашим информацијама, пре неколико дана у парку су били паркирани камиони ЕДБ који су највероватније оштетили тротоар.

Фреквенција: Елемент оквира којим се назначује колико се често врши оштећивање.

У наредна три века Турци су два пута тешко оштетили манастир.

Време: Овај ЕО идентификује време када је Агенс нанео штету Пацијенсу.
Семантички тип: време

Недавно је започела реконструкција те зграде која је прошле године оштећена пожаром, а „архитекта је оценио да нема опасности од рушења”, рекао је градоначелник.

Слика 4: Оштећивање (4).

2.2. Изазвати_емоцију

Оквир Изазвати_емоцију повезан је са оба глагола. У овој ситуацији АГЕНС дела и својим чином изазива интензивну негативну емоцију код ДОЖИВЉАВАЧА: *Ти си ме сломио. Уништио си ме!* и *Ја не желим тај брак – не желим да оштетим драгог малог Хенрија*. Централни елементи су АГЕНС и ДОЖИВЉАВАЧ, који представљају људе, а периферне елементе нисмо пронашли. Оштећивање, као и уништавање, типично је конкретно, физичко, а овде се смешта у емоционални или психолошки домен, те је активна метафора ПСИХИЧКО/ЕМОЦИОНАЛНО СТАЊЕ ЈЕ ФИЗИЧКО СТАЊЕ (в. Klikovac 2004, 178).

Сублексема *уништити* знатно чешће призива оквир Изазвати_емоцију у односу на сублексема *оштетити*.

Изазвати емоцију

Дефиниција:

Агенс дела и својим чином изазива одређену емоцију код **Доживљавача**.

Ти си ме сломио. **Уништио си ме!** **DN!**

Ја не желим тај брак – не желим да **оштетим** **драгог малог Хенрија** – али прилике су ми наметнуле ту помисао; а кад би господин Најтли стварно желео да се ожени, не бисте ваљда хтели да се уздржи од тога због Хенрија, шестогодишњег дечака, који о свему томе ништа не зна?

Централни елементи:

Агенс: **Агенс** је особа чији поступци узрокују емоцију код **Доживљавача**.
Семантички тип: свестан

Ти си ме сломио. **Уништио си ме!** **DN!**

Доживљавач: **Доживљавач** осећа емоцију коју је проузроковао Агенс.

Ти си ме сломио. **Уништио си ме!**

Слика 5: Изазвати_емоцију.

2.3. Стимулисати_емоцију

Оквир Стимулисати_емоцију у нашем корпусу повезује се само са сублексемом *уништити*. Ситуација уништавања у овом случају подразумева постојање неког феномена, названог СТИМУЛУС, који буди снажну, негативну емоцију код ДОЖИВЉАВАЧА, свесног ентитета који емоционално или психолошки реагује у односу на тај СТИМУЛУС. Разлика у односу на изазивање емоције састоји се у томе што у овом оквиру радњу врши догађај или неки други ентитет, док у претходном оквиру то чини човек.

Стимулисање емоције је „разрађеније” у односу на изазивање, те су забележена и два периферна елемента: СТЕПЕН (*потпуно*) и НАЧИН (*емотивно*).

Стимулисати емоцију

Дефиниција:

Неки феномен (**Стимулус**) буди одређену емоцију код **Доживљавача**.

Тај тренутак када сам сазнао шта је урадила, уништио ме је. Речима тај осећај не може да се опише.

Централни елементи:

Доживљавач: **Доживљавач** емоционално или психолошки реагује у односу на **Стимулус**.
Семантички тип: свестан

ТРАУМЕ СЛАВНОГ ФУДБАЛЕРА: Тај пораз **та је уништио**, морао да иде код психијатра

Стимулус: **Стимулус** је догађај или ентитет који доводи до одређеног емоционалног или психолошког стања **Доживљавача**.

Смрт брата ме је **потпуно уништила**. И данас осећам велику празнину.

Периферни елементи:

Степен: Означава се степен до којег **Стимулус** изазива емоцију код **Доживљавача**.
Семантички тип: степен

Смрт брата ме је **потпуно уништила**. И данас осећам велику празнину.

Начин: Начин на који **Стимулус** утиче на **Доживљавача**.

Увек је у питању нека топла и дивна прича која ме **емотивно „уништи”**.

Слика 6: Стимулисати_емоцију.

2.4. Ускраћивање

Глагол *оштетити* призива и оквир Ускраћивање. Ситуација означена овом сублексемом подразумева да АГЕНС утиче на ПАЦИЈЕНСА (нпр. *грађани, Србија, каријера*) тако да остане лишен СТАЊА_СТВАРИ, које би требало да има, потребно му је или га жели, и на тај начин му наноси штету. СТАЊЕ_СТВАРИ је елемент који је у нашем корпусу врло чест и изражен је именичком јединицом у ППК за + акузатив¹⁶, а под њиме се најчешће подразумева новац (у домену финансија) или потенцијални поени/голови (у спорту). Ипак, он не мора увек да буде исказан, те може да се односи на одређено право ПАЦИЈЕНСА да нешто поседује. Другим речима, иако није сасвим исто *оштетити државу* и *оштетити државу за милион динара*, често се

¹⁶ Употребу ове конструкције, која је карактеристична за лексему *оштетити* (и то за њен најчешћи колокат у савременом корпусу – *буџет*), ниједан описни речник српског језика није забележио.

подразумева да је и у првом случају оштећивање финансијске природе.

Ускраћивање

Дефиниција:

Овај оквир односи се на ситуацију у којој **Агенс** утиче на **Пацијенса** тако да **Пацијенс** остане лишен **Стања ствари**, које се очекује да има, које му је потребно или које жели. **Стање ствари** не мора увек да буде спецификовано и може да се односи на право **Пацијенса** да нешто поседује.

У прошле две године, српска полиција је у сарадњи са надлежним тужилаштвима открила **више десетина криминалних група које су оштетиле државу за више стотина милиона евра**.

Судије **су нас** драстично **оштетиле**, али део одговорности за пораз је на нама.

Семантички тип: догађај

Централни елементи:

Агенс: **Агенс** је особа која чини да **Пацијенс** остане без **Стања ствари**.

Семантички тип: разуман

Руководилац службе за продају лож-уља у једној фирми у САД и **службеник на рачунару** једне од највећих националних рафинерија нафте **оштетили су** ту рафинерију за 45 милиона литара горива (четири милиона долара) тако што су путем компјутера бележили мању продају од стварне, а разлику продавали приватно (крађа је, иначе, трајала шест година док није откривена!).

Пацијенс: **Пацијенс** је ентитет ком је ускраћено да има одређено **Стање ствари**.

Неиспуњавањем уговорних обавеза Живановић је **пословне партнере** **оштетио** за поменути износ.

Стање ствари: У питању је ситуација, стање ствари или стање поседовања које се **Пацијенсу** ускраћује.

Два пута смо погодили стативу, једном пречку, а онда нас је **судија** **оштетио** за два **пенала који су били чисти као суза**.

Слика 7: Ускраћивање (1).

Може бити исказано и СРЕДСТВО којим се врши ускраћивање (продајом робе на „црно”), НАЧИН на који се то чини (тако што су путем компјутера бележили мању продају од стварне, а разлику продавали приватно), као и ЦИЉ (да би ми „Макси“ изградио супермаркет), ПОСЕКОР (ваш) и ВРЕМЕ.

Периферни елементи:

Средство: Средство означава ентитет који **Агенс** користи како би утицао на **Пацијенса**.

Због сумње да је злоупотребила службени положај и продајом робе „на црно“ **оштетила** буџет Србије за 23,5 милиона динара, поднета кривична пријава против Н. А. из Кине, директорке предузећа „Асу“ из Панчева.

Начин: Искузује се начин на који **Агенс** ускраћује **Пацијенсу** **Стање ствари**.

Руководилац службе за продају лож-уља у једној фирми у САД и службеник на рачунару једне од највећих националних рафинерија нафте **оштетили** су ту рафинерију за 45 милиона литара горива (четири милиона долара) **тако што су** путем компјутера бележили мању продају од стварне, а разлику продавали приватно (крађа је, иначе, трајала шест година док није откривена!).

Циљ: Овај елемент оквира означава циљ због којег **Агенс** наноси штету **Пацијенсу** ускраћујући му неко право.

Семантички тип: стање_ствари

Ја бих желео, с обзиром да је тема корупција коју је он иницирао, да укажем на случај како су, ја сад цитирам један медиј, како **су** Кена и Шутановац **оштетили** Смедеревску Паланку **да би ми „Макси“ изградио супермаркет**.

Посесор: Ентитет ком **Пацијенс** припада.

Колико пута сте звиждали судији на фудбалској утакмици јер сте били убеђени да **је оштетио** **ваш** омиљени клуб?

Време: **Време** у ком се **Пацијенсу** ускраћује **Стање ствари**.

Они **су** **током 2002. и 2003. године** на основу фиктивног уговора са задругом о ангажовању радника **оштетили** „Жубор“ за 950.000 динара.

Слика 8: Ускраћивање (2).

2.5. Уништавање

У ситуацију уништавања укључена су два централна елемента: УНИШТАВАЧ односно УЗРОК и ПАЦИЈЕНС. Као и у случају оштећивања, УНИШТАВАЧ и УЗРОК се међусобно искључују; УНИШТАВАЧ је углавном свесна особа (човек или појам метонимијски повезан са њим), док је УЗРОК најчешће догађај или ентитет који заузима место у времену (*прокување, афера*), затим природне појаве/катастрофе (*ватра, ураган, цунами, експлозија*), предмети (*ракете*) или друштвени концепти (*сиромаштво, нацизам, технологија*).

У улози ПАЦИЈЕНСА може бити било који ентитет који не представља човека (*поверљиви списи, стабљике марихуане, две трећине манастирског комплекса*), а ситуација означена глаголом *уништити* подразумева да УНИШТАВАЧ или УЗРОК негативно утичу на ПАЦИЈЕНС, тако да изазову његов престанак постојања. ПАЦИЈЕНС могу представљати локације, грађевине и возила, друштвени

концепти¹⁷ (*каријера, живот, репутација, оптимизам*), информације (*документи, списи, докази*), медицински и здравствени ентитети (*фибросарком, меланом, бактерије*) и појмови из сфере природе и биома (*планета, шума, усеви*).

Уништавање

Дефиниција:

Уништавач/Разарач (свесни ентитет) или **Узрок** (догађај или ентитет укључен у такав догађај) негативно утичу на **Пацијенс**, тако да **Пацијенс** више не постоји.

Подофицер је био писар батаљона у повлачењу, управо је био примио наређење да **уништи** поверљиве списе, па је спаљивао све што му је долазило под руку пре него што ће кренути даље.

Албанска полиција је саопштила да је **уништила** 7.300 стабљика марихуане на северу земље.

Према првим проценама, **пожар у Хиљандару** **уништио** је две **трешине** манастирског комплекса.

Централни елементи:

Уништавач/Разарач: Свесни ентитет (најчешће човек) који врши намерну радњу која доводи до уништења **Пацијенса**.
Семантички тип: свестан, разуман

Када је постало јасно да Ейрон пропада, и када су на спену ступили владини истражитељи, **Андерсон** је **уништио** део компромитујућег материјала, мада постоји могућност да се део избрисаних архива поврати са рачунарских дискова.

Узрок: Ентитет или догађај који је одговоран за уништење **Пацијенса**.
Искључује се: Уништавач/Разарач

И **зрак** је **уништио** многе пупољке, па смо вишеу на пијаци плаћали 2,5, купину 1,5 евра за килограм, дупло више од досадашњег рекорда.

Пацијенс: Ентитет који бива уништен од стране **Уништавача/Разарача** односно **Узрока**.

Два третмана су била довољна да се у 88 одсто случајева потпуно **уништи** **фибросарком**, а меланом је био уништен у 17 одсто случајева. **CNI**

Слика 9: Уништавање (1).

Периферни елементи оквира (сл. 10–11) слични су онима који су посведочени у оквиру Оштећивања. И у овом случају може бити исказан САДРЖАВАЈУЋИ_ДОГАЂАЈ, у оквиру ког се одвија процес уништавања (у *терористичком нападу на Њујорк*). Затим следе СТЕПЕН (*потпуно*), ОБЈАШЊЕЊЕ (*јер ракете нису могле да буду испале*), ФРЕКВЕНЦИЈА (*често*), ИНСТРУМЕНТ (*пластичним експлозивом*), СРЕДСТВО (*бомбардовањем*), НАЧИН (*брутално*), као и МЕСТО (у *Ужичу*) и ВРЕМЕ (*дан раније*).

¹⁷ С. Иванова и С. Медведева (2023, 107) тврде да сви глаголи групе 'to destroy' колоцирају са лексемама које представљају 'вредности' (нпр. *слобода*) и 'претње' (нпр. *ропство*) (мада постоје и неутрални колокати), те уништавање вредности изазива негативна осећања и асоцијације, док се уништавање претњи перципира као позитиван чин. То важи и за глагол *уништити*.

Периферни елементи:

Садржавајући догађај: Овај елемент оквира означава догађај који траје или стање ствари које постоји у времену и обухвата период у коме се одвијају догађаји или стања ствари изражени циљном речју.

Градоначелник је најавио да ће наставити свој амбициозни програм о становању, образовању и изградњи Светског трговинског центра, који је уништен **терористичком нападом на Њујорк** 11. септембра 2001. године. **CNI**

Степен: Степен до којег је уништење извршено.
Семантички тип: степен

Пожар, за који се сумња да је подметнут, **потпуно је уништио** аутомобил а као колатерална штета страдало је још једно возило, „даву матиз“, у власништву Антићевих комшија.

Објашњење: Стање ствари на које **Уништавач/Разарач** реагује у процесу уништавања.
Семантички тип: стање_ствари

Дан раније град **је уништио** засаде **јер ракете нису могле да буду испалеене**.

Фреквенција: Означава се колико често **Уништавач/Разарач** односно **Угроза** уништава **Пацијенса**.

За поучу другим дечацима и њиховим родитељима, чија јуриња за првом гомилом свра **често уништи** каријере великим талентима.

Инструмент: Ентитет којим управља **Уништавач/Разарач**, који ступа у интеракцију са **Пацијенсом** да би извршио уништавање.
Семантички тип: физички_ентитет

Маскирана група екерата отерала је обезбеђење, упозорила станаре оближње зграде на опасност, а потом **је пластичним експлозивом уништила** опрему фирме која је као подизвођач радова била ангажована на асфалтирању главних ауто-саобраћајница у Србији.

Средство: Намерна радња коју је извршио **Уништавач/Разарач**, а којом се остварује уништавање.
Семантички тип: стање_ствари

У Вишеграду сам у почетку наишао на другачији пријем што не чуди с обзиром на то да **је НАТО бомбардовањем уништио** све мостове (на срећу, не и онај мост) а и електрана им је била тешко оштећена.

Напомена: Сваки опис радње уништавања која није обухваћена специфичнијим ЕО, укључујући епистемичке модификације (*вероватно*), силу (*тешко*, *слабо*), секундарне ефекте (*тихо*, *гласно*) и опште описе који упоређују догађаје (*на исти начин*). Може

Слика 10: Уништавање (2).

Осим тога, може бити истакнут МАЛЕФИЦИЈАР (*најугроженијим општинама у Банату = којима*). Њему је сличан ПОСЕКОР, ентитет ком ПАЦИЈЕНС припада (*наша производња*). Оквир садржи и елемент ЦИЉ (*да би доказао да је завршио*).

Најзад, у нашем корпусу је забележен и елемент УЛОГЕ, која представља категорију у оквиру које је ПАЦИЈЕНС уништен и изражена је конструкцијом са речју као (*као толерантног, динамичног интернационалног града*).

указивати на истакнуте карактеристике **Уништавача/Разарача** које такође утичу на радњу (*дрско, хладно, намерно, жељно, пажљиво*).
Семантички тип: начин

Где су данас остаци Жрнова, за који више од половина Београђана чак не зна ни да је постојао? Давне 1934. године краљ Александар Карађорђевић, **брутално** га је уништио, сравнивши га са земљом, заувек га пославши у заборав.

Малефицијар: Ентитет који, такође, посредно трпи уништавање.

Институт за кукуруз „Земун Поље” помоћи ће **најугроженијим општинама у Банату, којима** је поплава уништила усеве.

Циљ: Овај елемент оквира означава циљ због којег **Агенс** уништава **Пацијенс**.
Семантички тип: стање_ствари

Појасните које паре, која количина нафте да се запали Савезна скупштина, колико је тај уништио гласачких листића да би доказао да је завршио, да је способан да добије диплому.

Место: Овим ЕО идентификује се место на ком **Уништавач/Разарач** уништава **Пацијенс**.
Семантички тип: лоцираност_релативна

У Ужицу је уништена документација о Шљивовици.

Време: Овај ЕО идентификује време када је **Уништавач/Разарач** уништио **Пацијенс**.
Семантички тип: време

Дан раније град је уништио засаде јер ракете нису могле да буду испале.

Улога: Категорија у оквиру које је **Пацијенс** уништен.

Нећу дозволити да битанге и криминалци који желе да се понашају онако како смо видели у последњих 48 сати униште репутацију Сиднеја као толерантног, динамичног интернационалног града.

Посесор: Ентитет ком **Пацијенс** припада.

Циљ овог закона је да уништи нашу производњу, а да кренемо са увозом цигарета јер смо ми велики потрошачи тих чудних страних цигарета.

Слика 11: Уништавање (3).

2.6. Убијање

Оквир Убијање призива сублексема *уништити* када означава ситуацију у којој УБИЦА или УЗРОК доводе до смрти ЖРТВЕ. Напоменућемо да је овај оквир врло сличан претходном и да није лако повући прецизну границу међу њима, посебно ако се узме у обзир да у анализираним примерима није било много оних у којима сублексема *уништити* призива овај оквир. Ипак, основну разлику у односу на Уништавање представља постојање ЖРТВЕ, живог бића које умире као последица убијања. Затим, уместо УНИШТАВАЧА, смрт ЖРТВЕ изазива УБИЦА, који увек представља живо биће. У односу на Уништавање, прави се разлика између престанка постојања неживог ентитета и живог бића. Чини се, међутим, да глагол *уништити* типично не призива оквир Убијања, те би се потпунији опис оквира добила када би се у анализу укључили и други глаголи, попут лексема *искоренити*, *истребити*, *побити*, *поубијати*, *убити* итд.

Убијање

Дефиниција:

Убица или **Узрок** узрокује смрт **Жртве**.

Такође му је провокативна и теза о суштински геноцидној природи „хомо сапијенса“ који је, пре свега, уништио све друге људске врсте на Земљи, јер ми нисмо били једина.

Централни елементи:

Узрок: Неживи ентитет или процес који доводи до смрти **Жртве**.

Искључује се: Убица

Убица: Особа или свестан ентитет који узрокује смрт **Жртве**.

Искључује се: Узрок

Грчка митологија каже да је Атлантида била држава чији су становници били корумпирани похлепом и жељом за владањем да их је **Зевс** уништио.

Жртва: Живо биће које умире као резултат убијања.

Семантички тип: свестан, разуман

Градским властима дат је рок до 6. фебруара да **униште** 150.000 патака, док би једна приватна фирма требало да преради 60.000.

Слика 12: Убијање.

2.7. Победити_противника

Следи оквир у ком се описује ситуација када ПОБЕДНИК остварује победу над ГУБИТНИКОМ. Илустрована је експресивна употреба глагола *уништити*, која се углавном везује за домен спорта или политике. Наиме, спортска надметања и политичке расправе, које често могу бити агресивне, овде се разумевају као борбе или ратови (Klikovac 2000, 268; Klikovac 2004, 157). Изузетно убедљива, силовита победа концептуализује се као демолирање, разарање, те бисмо могли реконструисати појмовне метафоре УТАКМИЦА/ТАКМИЧЕЊЕ ЈЕ БОРБА и ПОБЕДА ЈЕ УНИШТЕЊЕ.

Уколико је уништавање усмерено ка непријатељу и води до победе, оно има позитивну форму (Ivanova and Medvedeva 2023, 112), за разлику од већине других оквира, у којима је уништавање негативно.

Од периферних елемената, може се јавити ТАКМИЧЕЊЕ (у *дербију*), затим СКОР, квантитативни елемент којим се мери прогрес током игре (*са четири гола*), као и СРЕДСТВО (*тројкама*) и ВРЕМЕ и МЕСТО ОДРЖАВАЊА.

Постоји јасна веза између физичког уништавања (изворног домена) и остваривања убедљиве победе (циљног домена), те се улога УНИШТАВАЧА пресликава на ПОБЕДНИКА, а ПАЦИЈЕНС је ГУБИТНИК. Средство уништавања би била тактика, начин игре и сл., степен уништења односи се на убедљивост победе, место уништавања је место такмичења, док је време уништавања повезано са временом одржавања меча.

Победити противника

Дефиниција:

Овом оквиру припадају речи које означавају ситуацију када **Победник** остварује победу над **Губитником**.

Брентфорд **уништио** Манчестер јунајтед **СА четири гола**.

Централни елементи:

Победник: Индивидуа или индивидуе које побеђују **Губитника**.

Никад не отписуј срце шампиона! **Новак** се „вратио из мртвих“, губио 2:0 у сетовима, па **уништио** Синера

Губитник: Индивидуа или индивидуе које побеђује **Победник**.

Након резултата локалних избора на Косову, лист наводи: „Дебакл: Вучић **уништио** Куртију”

Периферни елементи:

Такмичење: Догађај у ком се такмиче **Победник** и **Губитник**.

Ливерпул **уништио** Атлетико **у дербију**

Скор: Квантитативни елемент којим се мери прогрес током игре.

Брентфорд **уништио** Манчестер јунајтед **СА четири гола**

Средство: Акција **Победника** или **Губитника** која омогућава да се догоди пораз.
Семантички тип: стање_ствари

ЗВЕЗДА РАЗБИЈЕНА У ПРЕПУНОМ ПИОНИРУ: Уникс **ТРОЈКАМА** **уништио** црвено-беле, шутерска клиника Руса!

Време: Односи се на то када је **Победник** остварио победу над **Губитником**.

НИКОЛА ЈОКИЋ **УНИШТИО** ЛЕЈКЕРСЕ **СИНОТ!**

Место одржавања: Овај елемент оквира односи се на именовани простор одржавања такмичења, нарочито када је место изграђено са наменом да се на њему одржи такмичење.

Звезда **уништила** Бајерн **у Арени**! 101–77.

Слика 13: Победити_противника.

2.8. Довести_у_тешко_стање

Оквир Довести_у_тешко_стање сличан је оквирима Оштећивање и Уништавање, али је ипак посебно издвојен. Најпре, праћене су дефиниције лексеме *уништити* у Матичиним речницима, у којима се прави јасна дистинкција између престанка постојања и довођења у тешко стање. Довођење у тешко стање је заправо стање на рубу егзистенције, где један ентитет постоји¹⁸ у неком облику, али је потпуно изгубио изглед и функцију. Уколико је у питању предмет, у овом оквиру он више нема своја инхерентна својства, оно што га тим

¹⁸ У многим случајевима, чак и уз контекст, није лако одредити да ли ентитет постоји (бар у остацима) или је потпуно уништен.

предметом чини, и у том смислу као да више не постоји, попут *оргуља у Дому синдиката* (*Оргуље у Дому синдиката уништила је наша небрига*).

АГЕНС је човек или појам у метонимијској вези са њим, а УЗРОК углавном представља природне појаве/катастрофе (*поплаве, непогоде, ерупција*), догађаје (*ратови, санкције*), друштвене концепте (*одлуке, трагедија, политика*) и предмете (*меци, бомба*).

ПАЦИЈЕНС је повезан са друштвеним концептима (*живот, привреда, економија, буџет, држава*), са човеком и појмовима који су са њим повезани метонимијом (*Партизан, Звезда*), са грађевинама и местима, здравственим ентитетима (*бубрег, слезина*) и предметима (*оргуље, књиге*).

Довести у тешко стање

Дефиниција:

Агенс или **Узрок** утичу на **Пацијенса** тако да **Пацијенс** (или неки део на њему) заврши у потпуно неканонском стању, таквом да више не поседује првобитан изглед, функцију и сл. **Пацијенс** губи своје инхерентне карактеристике, готово потпуно је измењен и овакво стање **Пацијенса** је изузетно непожељно.

Такође, постоје **аеродроми који**, нажалост, нису у употреби, јер **су 1999. године уништени бомбама**, али могу да се ставе у неку функцију, за шта би требала велика инвестиција и велики новац. **СН**

Оргуље у Дому синдиката уништила је наша небрига.

Руски судија који је уништио Звезду и Рому лепо се уклопио у новонасталу атмосферу, али изгледа да није узимао лекове пре утакмице Португал – Холандија.

Централни елементи:

Агенс: Свесни ентитет, обично особа, који врши намерну радњу која доводи до тога да **Пацијенс** заврши у потпуно неканонском стању.

Семантички тип: разуман, свестан

Искључује се: Узрок

Њена мисија је била да приведе правди **неподирљиве моћнике који су** одузели животе десетинама хиљада људи и **уништили** животе милионима осталих.

Узрок: Ентитет који чини да **Пацијенс** заврши у потпуно неканонском стању.

Искључује се: **Агенс**

У стечајној маси је и одмаралиште Југово, неколико стотина метара дунавске обале, коју **су уништили поплаве.**

Пацијенс: Ентитет који због активности **Агенса** или **Узрока** губи свој првобитни изглед, функцију и сл. завршавајући у неканонском стању.

Наш европски пут **је уништио наше образовање** и ако се овако настави за одређени период, после извесног времена биће још гора ситуација него што је сада.

Слика 14: Довести_у_тешко_стање (1).

Дакле, овде АГЕНС или УЗРОК утичу на ПАЦИЈЕНС тако да заврши у стању које је озбиљно, органски нарушено. Физички, и даље у неком облику постоји, али је угрожен, што се види у свим наведеним примерима (нпр. *А поменути бивши министар је Стару планину*

управо тим малим просецима уништио). Овде је активна метафора УЗРОКОВАЊЕ НЕГАТИВНЕ ПРОМЕНЕ ЈЕ ДЕСТРУКЦИЈА (Миљковић 2014, 150).

ИНСТРУМЕНТ (погрешним хемикалијама), СРЕДСТВО (управо тим малим просецима), СТЕПЕН (готово потпуно), МАЛЕФИЦИЈАР (му), ПОСЕКОР (мој), МЕСТО и ВРЕМЕ јесу периферни елементи које смо уочили претрагом корпуса.

Периферни елементи:

Инструмент: Ентитет којим управља **Агенс**, који ступа у интеракцију са **Пацијентом** да би га довео у тешко стање.

Семантички тип: физички_ентитет

У питању је 3.500 вредних предмета из 19. века који **су уништени погрешним хемикалијама** у току нестручне дезинсекције. **CNI**

Средство: Намерна радња коју је извршио **Агенс**, а којом се остварује активност довођења у тешко стање.

Семантички тип: стање_ствари

А поменути бивши министар **је** Стару планину **управо тим малим просецима уништио**, апсолутно је довео до ерозије и до деградације која не може да се заустави, а верујте да то може да се уради и са далеководима и са водовима.

Степен: Степен у ком **Пацијент** завршава у неканонском стању.

Семантички тип: степен

Бомба велике разорне моћи постављена испод задњег дела возила, **готово потпуно је уништила** аутомобил.

Малефицијар: Ентитет који, такође, посредно трпи штету.

Он каже да **му је** ова трагедија **уништила** живот, али не верујем да му је уништила и карактер.

Посекор: Ентитет ком **Пацијент** припада.

Начелник одељења за странце у Љубљани Марјан Компара **је** у својој канцеларији буквално маказама разрецао, **уништио** **мој** пасош и рекао да сада идем у Македонију по документе.

Место: Место на ком се налази **Пацијент**, који трпи оштећивање.

Семантички тип: лоцираност_релативна

Начелник одељења за странце у Љубљани Марјан Компара **је у својој канцеларији** буквално маказама разрецао, **уништио** **мој** пасош и рекао да сада идем у Македонију по документе.

Време: Овај ЕО идентификује време када је **Агенс** нанео штету **Пацијенту**.

Семантички тип: време

Весић је у пријави навео да **је** „**током ноћи између 12. и 13. јула** група непознатих лица спрејевима **уништила** управо реконструисану фасаду зграде у Поп Лукиној улици бр. 12 на општини Стари град у Београду”.

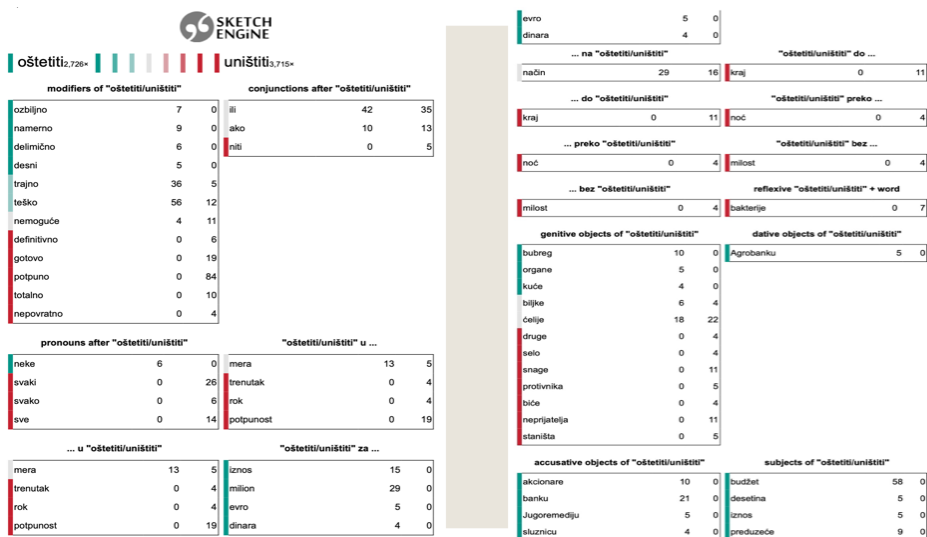
Слика 15: Довести_у_тешко_стање (2).

3. Поређење лексема оштетити и уништити

На платформи *Sketch Engine* могуће је преко алата *Word Sketch* генерисати скице речи, које сажето приказују понашања изабраних лексема, а уз помоћ алата *Word Sketch Difference* скице двеју лексема

Колокације, колигације, семантичке прозодије и семантичке преференције сматрају се важним аспектима анализе лексичког значења (в. Ноеу 2005, Louw 1993, Sinclair 1996, Stubbs 2002), посебно код проучавања блискозначница. Склоност лексема да се редовно повезују са одређеним колокатима представља и њихову преференцију да буду спојени са читавом семантичком класом сродних речи, што може представљати основу за генерализације, а Фрејмнет иде и корак даље од анализе колоката јер се фокусира и на семантичке улоге (Atkins et al. 2003b, 341).

202



Слика 19: Упоредна скица понашања.

4. Закључак

Анализом заснованом на теорији семантике оквира установили смо да се глагол *оштетити* повезује са Оштећивањем, Ускраћивањем и Изазивањем_емоције, а да глагол *уништити* призива оквири Уништавање, Довести_у_тешко_стање, Победити_противника, Убијање, Стимулисање_емоције и Изазивање_емоције. Сprovedено истраживање показује да је могуће подробно анализирати синтаксичко и семантичко понашање ових лексема у оквиру теорије семантике оквира. Премда се на основу анализе два глагола не могу донети генерализовани закључци о оквирима, чини се да њихов опис одговара корпусним подацима. Штавише, одабрани приступ омогућио је да се уоче значења која, на пример, описни речници не бележе (в. оквир Ускраћивање), као и да се идентификују елементи важни за опис значења, уз повезивање са синтаксичким конструкцијама (нпр., САДРЖАВАЈУЋИ_ДОГАЂАЈ јавља се само у пасивним клаузама и типично је реализован предлошко-падежном конструкцијом у + локатив).

Ситуација означена глаголом *оштетити* у најширем смислу подразумева наношење штете ПАЦИЈЕНСУ тако да заврши у лошем, неканонском стању. Штета је најчешће физичка, али може бити и финансијска и емоционална, као и шира – ускраћивање права. Оштећивање означава слабији степен деструкције, те се типично

оштећују мањи ентитети или делови већих, док је уништавање силовитије и типично се уништава цео ентитет. Исто тако, оштећивање се пре свега повезује са конкретним колокатима (артефакти, људско тело и здравље, биљке) и ређе се оштећује нешто што је у сфери апстрактног (домен финансија и спорта).

Са друге стране, лексема *уништити* чешће се користи да означи уништавање апстрактних концепата. Призвани оквири указују на то да ова лексема такође означава наношење штете, а екстремна штета доводи до престанка постојања (што је типично уништавање) односно убијања. Довођење у тешко стање, у ком ПАЦИЈЕНС губи природне карактеристике, али физички и даље постоји, налази се на континууму између типичног оштећивања и уништавања. Ова реч се користи и у домену спорта и политике, када се победа изједначава са уништавањем противника, као и у домену емоција и психе.

Синтаксичко понашање анализираних лексема је слично. Готово сви оквири садрже два обавезна елемента (АГЕНС/УЗРОК и ПАЦИЈЕНС, односно граматички субјекат и прави објекат). Изузетак представља употреба конструкције *за + акузатив* (елемент СТАЊЕ_СТВАРИ), која је карактеристична за лексему *оштетити*. Мање разлике уочавају се у сваком од оквира, и везују се за периферне елементе.

Ово истраживање, усмерено на глаголе *оштетити* и *уништити*, представља покушај да се приближи употреба ових лексема, и први је корак ка широј, будућој анализи глагола оштећивања и уништавања из угла теорије семантике оквира. Оваква анализа омогућава повезивање синтаксичке и семантичке структуре глагола, што је важно за разумевање њихове употребе у језику, а од посебног значаја може бити за учење српског као страног језика, за превођење, као и у контрастивним и типолошким проучавањима језика. Добијени подаци су применљиви и у лексикографији и домену обраде природних језика.

Важно је рећи и да спроведена анализа није у потпуности заокружена без, нпр., приказа синтаксичке и семантичке валенце глагола и описа других, сличних лексема које такође призивају описане оквири. Наредно истраживање би, уз употребу адекватних језичких ресурса, представило и те аспекте.

Литература

- [1] Atkins, B. T. S. (1994). Analyzing the verbs of seeing: A frame semantics approach to corpus lexicography. *Annual Meeting of the Berkeley Linguistics Society*, 20, 42–56.
- [2] Atkins, B. T. S., Fillmore, C. J., & Johnson, C. R. (2003a). Lexicographic relevance: Selecting information from corpus evidence. *International Journal of Lexicography*, 16(3), 251–280.
- [3] Atkins, B. T. S., Rundell, M., & Sato, H. (2003b). The contribution of FrameNet to practical lexicography. *International Journal of Lexicography*, 16(3), 333–357.
- [4] Fillmore, C. J. (1969). Verbs of judging: An exercise in semantic description. *Paper in Linguistics*, 1(1), 91–117.
- [5] Fillmore, C. J. (1982). Frame semantics. In *Linguistics in the Morning Calm* (pp. 111–137). Seoul: Hanshin Publishing Co.
- [6] Fillmore, C. J., & Atkins, B. T. S. (2000). Describing polysemy: The case of ‘crawl’. In Y. Ravin & C. Lacock (Eds.), *Polysemy* (pp. 91–110). Oxford University Press.
- [7] Fillmore, C. J., Johnson, C. R., & Petruck, M. R. L. (2003). Background to FrameNet. *International Journal of Lexicography*, 16(3), 235–250.
- [8] Hlebec, B. (2023). Defining English and Serbian verbs of damage. *Philologia*, 21(1), 1–21.
- [9] Hoey, M. (2005). *Lexical Priming: A New Theory of Words and Language*. Routledge.
- [10] Ivanova, S. V., & Medvedeva, S. N. (2023). Corpus analysis as a means of meaning disambiguation: Collocational analysis of near-synonyms of the verb ‘destroy’. *Professional Discourse & Communication*, 5(3), 107–125.
- [11] Klikovac, D. (2000). *Semantika predloga: Studija iz kognitivne lingvistike*. Beograd: Filološki fakultet.
- [12] Klikovac, D. (2004). *Metafore u mišljenju i jeziku*. Beograd: Biblioteka XX vek; Knjižara Krug.
- [13] Louw, B. (1993). Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In M. Baker, G. Francis, & E. Tognini-Bonelli (Eds.), *Text and Technology* (pp. 157–176). Amsterdam: John Benjamins.
- [14] Rasulić, K., & Klikovac, D. (2014). *Jezik i saznanje: hrestomatija iz kognitivne lingvistike*. Univerzitet u Beogradu: Filološki fakultet.

- [15] Ruppenhofer, J., Ellsworth, M., Petruck, M. S., Johnson, C. R., & Scheffczyk, J. (2016). *FrameNet II: Extended theory and practice (revised November 1, 2016)*. https://framenet.icsi.berkeley.edu/fndrupal/the_book
- [16] Sinclair, J. (1996). The search for units of meaning. *TEXTUS*, 9(1), 75–106.
- [17] Song, Q. (2021). Effectiveness of Corpus in Distinguishing Two Near-Synonymous Verbs: ‘Damage’ and ‘Destroy’. *English Language Teaching*, 14(7), 8–20.
- [18] Stubbs, M. (2002). Two quantitative methods of studying phraseology in English. *International Journal of Corpus Linguistics*, 7(2), 215–244.

- [19] Марковић, А. и др. (2021). Лексичка база Фрејмнет: неколико примера оквира из домена ризика. *Infotheca – Journal for Digital Humanities*, 21(1a), 7–36.
- [20] Миљковић, В. (2014). Семантика глагола *уништити* у светлу теорије менталних простора. У М. Ковачевић и Ј. Петковић (ур.), *Савремена проучавања језика и књижевности* (V/1, стр. 139–151). Крагујевац: Филолошко-уметнички факултет.
- [21] Поповић, Јб. (1997). Интегрални модел међуодноса синтаксичког система, речника и текстова. *Научни састанак слависта у Вукове дане*, 26(2), 469–482.
- [22] Поповић, Јб. (2003). Интегрални реченични модели и њихов значај за лингвистички опис и анализу корпуса. *Научни састанак слависта у Вукове дане*, 31(1), 201–220.
- [23] Стипчевић, Б. (2010). Глаголи с рекцијском допуном у беспредлошком инструменталу и њихови реченични модели. *Научни састанак слависта у Вукове дане*, 39(3), 175–254.

Syntactic and Semantic Analysis of the Verbs *ošteti* and *uništi* from the Perspective of Frame Semantics Theory

Marina Bagi

Summary

This paper analyzes the verbs *ošteti* ('to damage') and *uništi* ('to destroy') from the perspective of the frame semantics theory. Starting from the assumption that syntactic and semantic phenomena are mutually correlated, the theoretical framework for this study is based on a cognitive-semantic approach developed by Charles Fillmore, which explicitly links the meaning of the word with the syntactic contexts in which it appears. The core thesis of the frame semantics theory is that the meanings of sublexemes should be viewed in relation to semantic frames – schematic depictions of conceptual structures and belief patterns underlying word meanings. This theory is fundamental to FrameNet, an electronic lexicographic database based on a corpus, within which the frames invoked by various lexemes are presented.

This study examines six hundred examples of the use of the verbal lexemes *ošteti* ('to damage') and *uništi* ('to destroy'), excerpted from the electronic *Corpus of Contemporary Serbian (SrpKor2021)*, with the aim of describing and examining their syntactic and semantic behaviour, as well as illustrating their usage. Consequently, the frames evoked by these verbs are described, along with the elements of the frames, and conclusions are drawn regarding the syntactic-semantic interface. Furthermore, since these verbs share similar semantics, the aim is to investigate both the similarities and differences between them using tools available on the *Sketch Engine* and *No Sketch Engine* platforms.

Keywords: frame semantics theory, syntax, semantics, verb *ošteti*, verb *uništi*, natural language processing, Serbian language.

Дигитализација српског књижевног наслеђа ијекавског изговора (1840–1920) при Центру за дигиталну хуманистику Филозофског факултета Пале (прва фаза)

Научни рад

DOI: 10.18485/judig.2025.1.ch12

Срђан Шућур¹,  0009-0003-6815-2887

Јелена Марковић²  0000-0003-4962-3754

Апстракт

У овом чланку представљамо пројекат Дигитализација српског књижевног наслеђа ијекавског изговора (1840-1920) при Центру за дигиталну хуманистику Филозофског факултета Пале (прва фаза), односно изградњу Корпуса српског књижевног наслеђа ијекавског изговора, који настаје као примарни резултат поменутог пројекта. У постојећим корпусима српског језика ијекавски изговор није представљен пропорционално својој заступљености код говорника српског језика као матерњег, а самим тим мање је доступан и за корпусна истраживања. Стога поменути Корпус репрезентативне књижевности писане на српском језику ијекавског изговора у изградњи, који иначе представља проширење корпуса SrpELTeC+ (из објективних разлога доминантно екавског изговора) у оквиру COST акције Distant Reading for European Literary History (COST Action CA16204), има за циљ обogaћивање и проширивање постојећих корпуса ијекавског изговора српског језика. Одредница „прва фаза“ у називу пројекта наговјештава да су планирана и будућа проширења и надоградња корпуса.

¹ Универзитет у Источном Сарајеву, Филозофски факултет Пале, Катедра за англистику, srdjan.sucur@ffuis.edu.ba

² Универзитет у Источном Сарајеву, Филозофски факултет Пале, Катедра за англистику

У тексту приказујемо актуелни пресјек реализације пројекта, односно процес формирања Корпуса. Представљамо одабир текстова за дигитализацију, рјешавање проблема достизања репрезентативности Корпуса, процес дигитализације одабраних текстова уз примјере рјешења, уз најаву публиковања припремљених дигитализованих облика на сајту Центра за дигиталну хуманистику при Филозофском факултету Пале, а у организацији Друштва за језичке ресурсе и технологије из Београда ЈеРТех.

Пројекат у настајању, односно прва фаза, обухвата приповијетке Петра Кочића, Светозара Ђоровића, Васе Кондића, Марка Поповића, Милана Трифуновића, Јоаникија Памучине, Михајла Милановића, Николе Видаковића, Радована Перовића Тунгуза, Луке Грђић-Бјелокосића, Милана Трифуновића, Ђорђа Чокорила, Милке Алексић Гргурове, Љубице Поповић, Олге Ракић и Јелице Беловићеве, потом путописе Константина Хаџиристића, Јове Бесаровића, Саве Косановића, Ристе Бесаровића, Саве Пјешчића и Марка Цара, те романе Светозара Ђоровића и Радована Перовића Тунгуза.

У предстојећем периоду планирано је проширивање Корпуса ијекавског изговора књижевном критиком, кратком причом, као и другим жанровима, у циљу формирања релативно прихватљивог обима дигитализованих ресурса ијекавског изговора српског језика из разлога лингвистичке и екстралингвистичке природе.

Кључне ријечи: SrpELTeC+, дигитализација, српско књижевно наслеђе, ијекавски изговор

1. Увод

Статус српског језика уопште у актуелном пресјеку стања у дигитализованим ресурсима може се окарактерисати на различите начине, у опсегу од релативно повољног, уважавајући критеријум обима и улоге заједнице која се њиме служи као матерњим, до релативно неповољног, по критеријуму поређења са значајнијим свјетским језицима и њиховим дигиталним ресурсима и напретком³. У сличним наводима и тумачењима, као норму и нехотично најчешће

³ Када је у питању изградња корпуса српског језика у дијахронији, вриједи истаћи два ранија слична подухвата дигитализације, у основи лексикографске природе. Први, реализован у оквиру пројекта назива Речник српског језика од 12. до 18. века, који је покренут 2013. године, за циљ има дигитализовање ћириличких рукописа (~10.000 страница) Гаврила Стефановића Венцловића (1680–1749), помоћу софтвера *Transkribus*, а у сврху прикупљања грађе за Историјски речник српског језика (Polomas, et al., 2023). Други подразумијева формирање корпуса славеносрпских текстова (1750–1848), намијењеног изради Славеносрпског речника (Милановић, 2017).

узимамо екавски изговор српског језика, што, свакако, има своје друштвено-политичке и социolingвистичке разлоге. Ијекавски изговор српског језика номинално представља изговор равноправан екавском⁴, али услови у којима српска говорна заједница живи, директно и индиректно, утичу на неповољан положај ијекавског изговора у процесима дигитализације српског језика.

Говорећи о статусу српског језика и књижевности у Републици Српској Бабић (2021, 74) истиче поновно актуелизовање проблема „екавице и ијекавице у српском језику – у корист екавице, с циљем језичког унифицирања Срба“, од стране одређених политичких кругова везаних за први сазив Скупштине Републике Српске. Истовремено подсјећа на *Закон о службеној употреби језика и писма у Републици Српској* (од 29. јуна 1996. године) и њиме прописану обавезу употребе екавског изговора⁵. Да се на овој обавези истрајавало, даље наглашава Бабић, били би поништени вијекови „српске ијекавске традиције“, а „одустајањем од ијекавице“⁶ и сама основица српског књижевног језика би се „врло лако прогласила туђом“⁷.

Једноставна претрага упита попут „ијекавски изговор српског језика“, „корпуси српског језика ијекавског изговора“, коришћењем ћириличног и латиничног писма, прилично је индикативна за статус ијекавског изговора у дигиталним изворима. Резултати таквих претрага углавном воде ка општим прегледима статуса јужнословенских језика који су објављени у Босни и Херцеговини или Хрватској, а најчешће повезују ијекавски изговор са језичким варијантама које се не одређују као српске. Међу резултатима налазимо и коментаре проф. др Душка Витаса, који без околишања каже: „ако изгубимо ијекавицу, други ће нам одузети књижевност“⁸, уз навод да се:

[...] у кругу хрватских рачунарских лингвиста разматра проблем разликовања српског и хрватског језика и граница која се аутоматски подвлачи јесте екавски и ијекавски говор, а ми

⁴ Предмет Паун Јовановић против Србије (Представка број 41395/15), пред Европским судом за људска права у Стразбуру слика нешто другачију екстралингвистичку реалност.

⁵ У јесен 1993. године „предност екавице у језичком стандарду“ дата је одлуком руководства у Палама, но то није подразумијевало литерарни и разговорни стил (Милосављевић, 2019: 46, 51).

⁶ Идеја о предности екавице у Републици Српској одбачена је 1998. године (Милосављевић, 2019: 46).

⁷ О покушају подвођења свега што је ијекавско под хрватско, те подвођењу писаца свих националних одређења, са цијелокупног простора БиХ, који су писали ијекавицом под „босански језик“ говори Стојановић (2010: 17, 20).

⁸ <https://www.dnevnik.rs/kultura/vitas-ako-izgubimo-ijekavicu-drugi-ce-nam-oduzeti-knizevnost-26-06-2017>

немамо добар узорак српских ијекавских говора на основу кога би било могуће конструисати моделе за програм који одлучује да ли је неки текст на српском или хрватском језику.

1.1. Корпус српског књижевног наслеђа ијекавског изговора (1840–1920) – основни подаци

Имајући у виду значајно кашњење научне јавности у формирању узорака и корпуса ијекавског изговора српског језика, као и недавно публикованих стотину романа српског језика (Krstev, 2021) у оквиру COST акције под називом „Distant Reading for European Literary History“ (Schöch, et al. 2021), односно успјех српске академске заједнице у погледу заузимања мјеста које јој и припада у тијелу романа европских језика у периоду 1840–1920, иницирана је идеја о проширивању поменутог корпуса ијекавским изговором, у фазама, користећи друге жанрове од значаја, попут приповијетки, путописа, књижевне критике, епистоларне форме, администативних и других жанрова. Поменута идеја у својој првој фази, односно првом сегменту, реализује се кроз пројекат под називом *Дигитализација српског књижевног наслеђа ијекавског изговора (1840–1920) при центру за Дигиталну хуманистику Филозофског факултета Пале (прва фаза)*, који је подржан од стране Министарства за научнотехнолошки развој и високо образовање Републике Српске.

Корпус који треба да буде резултат наведеног научноистраживачког пројекта представља проширење већ објављеног корпуса SrpELTeC⁹ (Stanković, Krstev & Vitas, 2024), започетог у оквиру завршене COST акције CA16204. Овај пројекат наставак је досадашње плодне сарадње чланова пројектног тима са Друштвом за језичке ресурсе и технологије JePTex из Београда¹⁰.

⁹ Формирању српске компоненте корпуса ELTeC посвећен је тематски број часописа *Инфотека* 21(2). У овом броју описан је настанак корпуса SrpELTeC „од празне листе, до колекције од стотину романа“ (Trtovac, Milnović & Krstev, 2021, 7–25). О метаподацима придруженим овом корпусу пише Крстев (Krstev, 2021, 26–42), док процес анотације описују Станковић, Крстев, Шандрих Тодоровић и Шкорић (Stanković, Krstev, Šandrih Todorović & Škorić, 2021, 43–59).

¹⁰ Досадашња сарадња читава се у формирању трокомпонентног корпуса КорСАНг, који, осим превода са српског на енглески и са енглеског на српски, чине и аргументативни састави србофоних студената англистике на матерњем српском језику (КорССАНг). За више података о корпусу видјети Радоња и Шућур (2021). Осим тога, сарадња се огледа и у значајној помоћи у изради двају докторских дисертација, од којих је једна већ успјешно одбрањена на Филозофском факултету Пале Универзитета у Источно Сарајеву (Шућур, 2022), а друга у фази израде.

Претходно споменути корпус SrpELTeC+ је, из разлога објективне природе, који подразумевају разне социолингвистичке факторе, те доступну грађу, доминантно екавског изговора. Његово проширење корпусом репрезентативне књижевности писане на српском језику ијекавског изговора омогућиће синхронно и дијахронно проучавање ових двају изговора српског језика, те ће представљати, између осталог, значајан искорак у приближавању потоњег изговора широј читалачкој публици, као и подстијек за будућа проширења електронских корпуса српског језика. Свакако још је важније да један од доприноса буде већа присутност ијекавског изговора српског језика у различитим облицима на глобалној мрежи, а самим тим и његово препознавање као једног од два равноправна изговора српског језика.

2. Одабир текстова за дигитализацију

Текстуални жанрови који су планирани у првој фази јесу романи, приповијетке и, путописи. Осим романа, који су свакако циљни жанр постојећег корпуса SrpELTeC+, друга два одабрана жанра сматрали смо комплементарним жанру романа имајући у виду да су у питању жанрови писаног језика, затим да су у питању текстови који по подобности за језичка, књижевна и шира друштвено-историјска проучавања могу бити од великог значаја.

Одабир текстова за дигитализацију повјерен је члановима пројектног тима са Катедре за компаративистику, Катедре за библиотекарство и Катедре за србистику Филозофског факултета Универзитета у Источном Сарајеву. Руководи се критеријума за израду корпуса ELTeC, у разматрање су узета како канонска, тако и мање позната или непозната дјела први пут објављена на српском језику ијекавског изговора, на ћирилици¹¹, у периоду од 1840–1920, те је провјерена њихова доступност у аналогном и дигиталном формату у библиотечким фондовима у земљи и окружењу.

Након што је формиран прелиминарни списак публикација које задовољавају претходно наведене критеријуме за израду корпуса, чланови тима су оцијенили значај појединачних публикација за корпус. Тиме је за дигитализацију одабрано 46 публикација, и то 1. *приповијетке*: Петра Кочића, Светозара Ђоровића, Васе Кондића, Марка Поповића, Милана Трифуновића, Јоаникија Памучине, Михајла Милановића, Николе Видаковића, Радована Перовића Тунгуза, Луке Грђић-Бјелокосића, Милана Трифуновића, Ђорђа

¹¹ Уз изнимку приповијетке *Vićo*, аутора Николе Видаковића, која је објављена на латиници.

Чокорила, Милке Алексић Гргурове¹², Љубице Поповић, Олге Ракић и Јелице Беловићеве, 2. *путописи*: Константина Хадиристића, Јове Бесаровића, Саве Косановића, Ристе Бесаровића, Саве Пјешчића и Марка Цара, и 3. *романи*: Светозара Ћоровића и Радована Перовића Тунгуза.¹³ Међу публикацијама налазе се индивидуална издања, збирке приповиједака, те приповијетке објављене у периодици¹⁴: *Босанска вила*, *Дабро-босански источник*, *Даница*, *Зора*, *Женски свет* и *Вошњак*. Заступљеност појединачних публикација приказана је на Графикону 1.



Графикон 1. заступљеност појединачних публикација у корпусу

3. Приповијетке

За дигитализацију су одабране следеће приповијетке Петра Кочића; *Гроб слатке душе*, *Мрачајски прото* и *Мргуда*, из збирке приповиједака *С планине и испод планине* (1907), те *Вуков гај*, *Змијање* и *Кроз мећаву* (*Јауци са Змијања*, 1910).

Светозар Ћоровић заступљен је приповијеткама *Мали просјак*, *Под вјештичиним кровом* и *У мајчину загрљају* (*Из моје домовине*,

¹² Ауторка је у бројевима *Босанске виле* у којима су објављена дјела одабрана за дигитализовање потписана као Милка Гргурова.

¹³ Осим публикација које су ангажовањем чланова тима у пројектним активностима укључене у Корпус, придружене су и двије већ дигитализоване публикације, захваљујући сарадњи са групом ЈеРТех из Београда.

¹⁴ Српску периодику у Босни и Херцеговини (пored *Босанске виле* и *Зоре* ту су и *Пријеглед* *мале библиотеке*, *Развитак* и *Српска омладина*) у вријеме аустроугарске власти истражује Ласица (2016, 2020).

1898), те приповијетком *Разорено гнијездо* (периодично објављивана у *Зори*, 1898, 1–9).

На коначни списак за дигитализовање уврштене су и приповијетке Васе Кондића; *Мајко ја идем, И шуме наше плачу, Сухо грање* (Приповијетке, 1910), те Марка Поповића; *Невјера*, затим *Мара, Мејрима, и Сеја* (Књижевни радови Марка С. Поповића – *Родољуба*¹⁵, 1893), као и приповијетка *Све за народ*¹⁶ (1894).

Првом фазом пројекта обухваћене су и приповијетке Милана Трифуновића; *Млади Станковићи, Раде, Чича Стојан* (Из равне Посавине, 1906) и *Општинска вага* (објављена у 11. броју *Женског света* 1906), затим приповијетке Михајла Милановића; *Јерко Урошевић, Стана, Христова икона* (Босанке, 1903), Николе Видаковића; *Вицо* (*Вошњак*, 1904), Радована Перовића Тунгуза, објављене у *Босанској вили*; *Има љубави* (1906, 1–2), *Леле Турци* (1909, 21–22), *Тек након пет година* (1910, 1–2) и *Амајлија над амајлијама* (1911, 3–12).

По једном приповијетком заступљени су Лука Грђић-Бјелокосић; *Крвнина*, (1892) и Ђорђе Чокорило; *Крај Кошеве* (Зора, 1901).

Као и приликом формирања корпуса SrpELТес, SrpELТес-ext и SrpELТес+, изазов је представљало проналажење женских аутора¹⁷, те је тако у ову фазу дигитализације уврштено 8 дјела 4 различита женска аутора. За дигитализацију је одабрано пет дјела Милке Алексић Гргурове¹⁸, која су претежно периодично објављивана у *Босанској вили*; *Двије сестре* (1897, 1–3), *Несретан случај* (1898, 14–18), *Преко љествица* (1900, 17), *Циганка* (1900, 1–3), *Три пријатеља* (1901, 13–

¹⁵ https://digitalna.nb.rs/view/URN:NBN:RS:ND_59067F874356CCC9C2256722F6BB18CC

¹⁶ https://digitalna.nb.rs/view/URN:NBN:RS:ND_D66BA3B89CF0FA0A3C2EB2F3509458CC

¹⁷ Током отоманске власти могућности за развој образовања и културе на Централном Балкану биле су изразито ограничене. У Босни и Херцеговини у том периоду прилику да се образују имали су углавном бегови и њихове супруге. Прилике постају нешто повољније након окупације ових крајева и њихове анексије од стране Хабзбуршке монархије, која је женама понудила више основних права од оних које су уживале током Османског царства. Најсеобухватнији приказ стања тадашње културе и образовања и њиховог значаја за жене дат је у сарајевском часопису *Српкиња* (1913) (Hawkesworth, 2000: 89, 243). Период од 1857. године (када је основана прва српска, женска, основна школа) до окупације обиљежио је просвјетни и хуманитарни рад Хаци Стаке Скендерове (*Босанска вила*, 1903, XVIII (23/24), 393–395), а упоредо и касније и велике српске добротворке, Енглескиње Мис Аделине Паулине Ирби (енг. *Miss Adeline Paulina Irby*) (*Босанска вила*, 1904, XIX (13/14), 269).

¹⁸ Куриозитет је да је прва приповијетка Милке Гргурове објављена у *Босанској вили* 1896. године (број 21), *Две жене*, писана екавицом, док су приповијетка *Двије сестре* (*Босанска вила*, 1897, бројеви 1–3), као и преостале приповијетке одабране за дигитализацију (у *Босанској вили* објављиване у периоду од 1897–1901), писане ијекавицом. Истовремено, њена приповијетка *У часу опасности*, објављена 1897. године у мостарској *Зори*, такође је писана екавицом.

14). По једном приповијетком у корпусу су заступљене Љубица Поповић; *Бјелка (Босанска вила, 1899, 1)*, Олга Ракић; *Провала (Босанска вила, 1899, 13–14)* и Јелица Беловићева; *Зима у Кључу (Босанска вила, 1907, 1)*.

4. Путописи

Дигитализацијом је обухваћено и пет путописа, међу којима су путопис Константина Хаџиристића *Из Босне о Босни*, уједно и најстарија публикација уврштена у прву фазу овог пројекта, периодично објављивана у *Даници* (1868, 21–34, 1869, 32–32), као и најобимнија публикација, путопис Марка Цара; *Венеција: успомене с пута* (1891). Преостала три путописа јесу *Из Стамбола у Сарајево (Босанска вила, 1886, 14)*, Јове Бесаровића, затим *Ријеч двије са мога путовања у Цариград* (1908), Ристе Бесаровића, те путопис Саве Пјешчића *Од Требиња до Мостара*, периодично објављиван у *Дабробосанском источнику* (1890, 10–12).

5. Романи

У корпус ће бити укључени и једно издање које је дигитализовано приликом формирања српске компоненте корпуса ELTeC; роман Радована Перовића Тунгуза *Из земље плача* (1906), те *Дивљак* (1913), Светозара Ћоровића¹⁹.

6. Дигитализација корпуса

Пројектом је предвиђено одржавање пет радионица, које су реализоване у периоду од маја до октобра 2024. године, у организацији Друштва за језичке ресурсе и технологије ЈеРТех. Предавач на радионицама била је проф. др Ранка Станковић, а учесници чланови пројектног тима. Током радионица чланови тима су, кроз презентације, консултације и практичне задатке упознати са изазовима дигитализације и обучени за њено спровођење. Почевши од општих тема у оквиру радионице под називом „Дигитализација: од слике до дигиталног текста“, затим преко радионица названих „Обрада дигиталних текстова“, „Аутоматска анотација текстова“, „Читање, читање изблиза [корпуси, конкорданце] и читање из далека [статистичка обрада]“, уприличено је и представљање могућности база података коришћењем алата ТХМ („Текстометријска анализа текста коришћењем алата ТХМ“).

¹⁹ https://digitalna.nb.rs/view/URN:NB:RS:ND_7C39AFC1F4EC2BD9E29F4BD666069A7F

Библиотеке које су у својим фондовима пронашле публикације одабране за дигитализацију и претходно их (или за потребе овог пројекта) скенирале, као и број појединачних публикација, наведени су у Табели 1. Три књиге преузете су са интернетског сервиса *Google Books*.

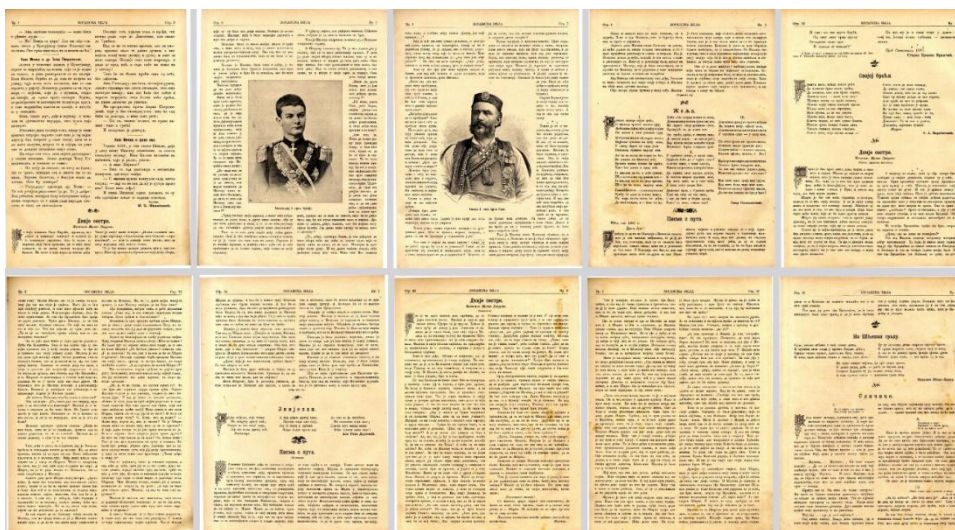
Табела 1. Библиотечки фондови из којих су преузете публикације

Фонд	Број публикација
Народна и универзитетска библиотека Републике Српске	19
Народна библиотека Србије	8
Библиотека Филозофског факултета Пале	3
Универзитетска библиотека "Светозар Марковић"	3
Народна библиотека "Радислав Никчевић" Јагодина	3
Библиотека Сарајева	3
Google Books	3
Национална и универзитетска библиотека Босне и Херцеговине	2
Библиотека Матице српске Нови Сад	1
Народна библиотека "Филип Вишњић" Бијељина	1
Укупно	46

Током одвијања првих трију радионица формирана је база највећег броја скенираних публикација и приступило се њиховој припреми за OCR (енг. *Optical Character Recognition*), тј. оптичко препознавање карактера, односно „рашчитавање“. Упоредо са формирањем базе ажурирани су метаподаци о публикацијама, и то: о првом издању, односно издању које ће бити дигитализовано, називу публикације, обиму, броју стубаца, жанру, аутору публикације, години рођења и смрти, периодици/збиркама приповиједака из којих су публикације/путописи издвајани, те институцијама које су помогле дигитализацију.

Припрема публикација за OCR подразумијевала је: 1) издвајање појединачних приповијетки из збирки приповиједака, 2) обједињавање приповијетки објављених периодично у јединствене документе.

Примјер приповијетке Милке Алексић Гргурове *Двије сестре* објављене у узастопна три броја *Босанске виле* (1897, 1–3) , обједињене у јединствен документ, дат је на Сlici 1.



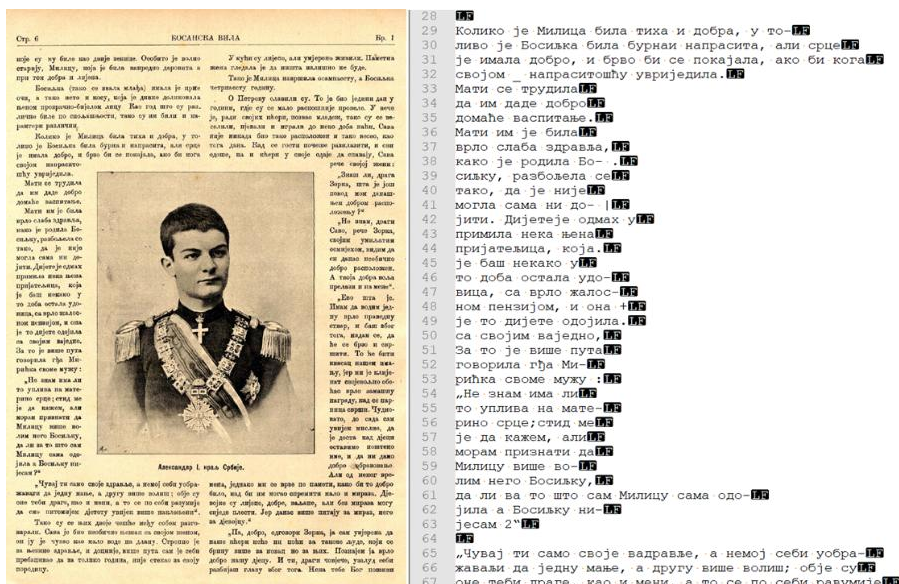
Слика 1. Приповијетка *Двије сестре*, изворно објављена у три узастопна броја *Босанске виле*, обједињена у један документ

По обављеној припреми публикација, OCR је вршен уз сарадњи са JePTex-ом, помоћу интернетског алата доступног на линку texase.jerteh.rs, заснованом на *Tesseract Open Source OCR Engine*-y²⁰.

Успјешност OCR-а била је условљена изворним стањем оригиналних публикација које су скениране, тј. факторима попут мрља, боје папира (последица старења), печата, поравнања при скенирању, биљешки накнадно писаних оловком и сл. На исход OCR-а утицала су и својства изворног текста, попут графичког уређења (нпр. централно постављене слике, в. Слика 2.) и форматирања (текст у два ступца, украси), као и накнадна обрада скенираних публикација, додавањем воденог жига (енг. *watermark*).

Након што су прикупљени OCR-овани документи, било их је неопходно додатно уредити, прије приступања аутоматској и ручној корекцији. То уређивање подразумијевало је уклањање „вишкова“; у случају публикација објављиваних периодично, било је неопходно уклонити дијелове текстова публикација које су у неком броју претходиле тексту одабраном за дигитализацију, или су га слиједиле, затим називе периодике, њихове бројеве, и сл. У приповијеткама издвојеним из збирки приповиједака уклањани су називи приповијетке/име аутора, који се у изворном тексту смјењују на парним, односно непарним страницама.

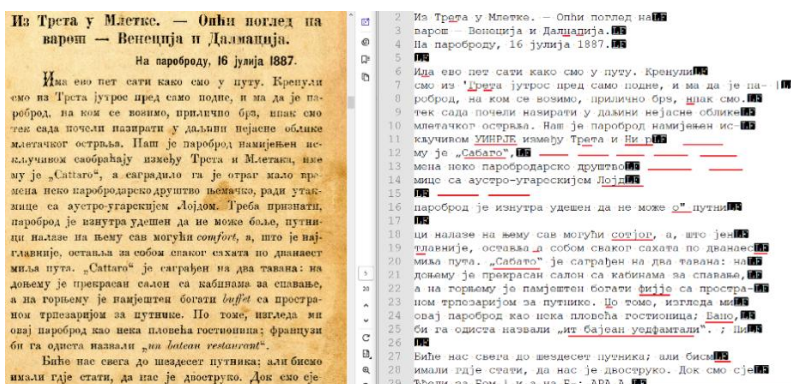
²⁰ <https://github.com/tesseract-ocr/tesseract>



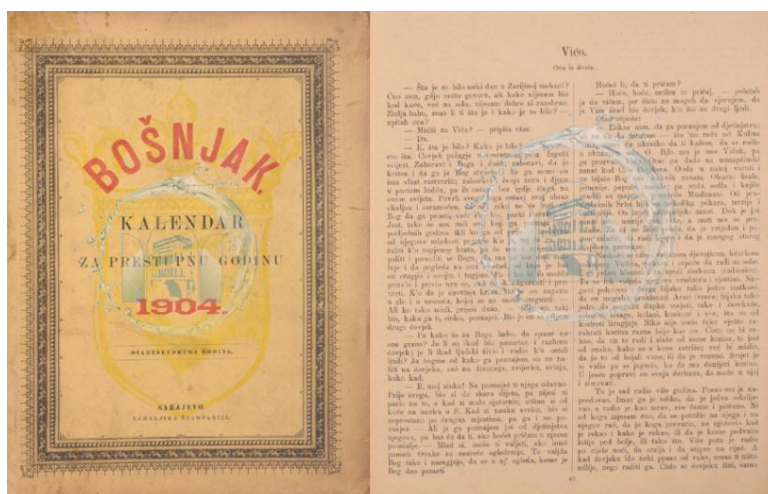
Слика 2. Примјер OCR-а текста објављеног у периодичи у два ступца са централно постављеном сликом

Услед претходно описаних фактора, у OCR-ованим документима, и поред аутоматске корекције, понекад је долазило до неразликовања ћириличних слова (м-л, т-г, и-н, м-н, п-н, в-з, с-е, ц-п), или интерпункцијских знакова (наводници, тачке, зарези, цртице и црте). У неким случајевима недостајале су ријечи (Слика 3.), или читави пасуси, а у случају једине публикације на латиници (приповијетка *Vićo*, в. Слика 4.), OCR због воденог жига није био успјешан, те је било неопходно прекуцати текст публикације. Повремено је долазило и до преклапања текста објављеног у два ступца.

Ручна корекција OCR-ованих текстова вршена је софтвером Notepad++, проширеним језицима ELTeC1 и ELTeC2, који су развијени за српски језик, помоћу електронских рјечника (Krstev, 2008), што омогућава аутоматску корекцију.



Слика 3. Примјер OCR-ованог изворног текста са недостајућим ријечима



Слика 4. Приповијетка *Vićo* и централно постављени водени жиг који је отежао OCR

Ручна корекција подразумијевала је и уређивање пасуса (тзв. „тврди крај реда“ CR LF), како би се подударали са оним у оригиналним публикацијама, те спајање ријечи које су у оригиналу прекинуте на крају реда. Паралелно са ручном корекцијом вршена је и основна анотација, тј. обиљежавање текста етикетама у складу са TEI (енг. *Text Encoding Initiative Guidelines*) (Bernard, Schöck & Odebrecht, 2021). То обиљежавање подразумијева означавање поглавља, фуснота, дијелова текста који су на страном језику, или су истакнути на неки начин (курзив, подебљана/велика слова и сл.). Коришћене етикете дате су у Табели 2.

Табела 2. TEI етикете коришћене при основној анотацији

Етикета	Означава
	почетак и крај поглавља
<head> </head>	наслов поглавља
<pb n="X" ²¹ /> <pb n="X.Y" ²² />	почетак нове стране из оригинала (убацује се у непрекинут ред)
<title> </title>	назив умјетничких и других дјела
<foreign> </foreign>	текст на страном језику
<hi> </hi>	истакнути дио текста
<quote> <l> </l> </quote>	стихове пјесама
<ref target="# NX" ²³ /> <div type="notes"> <note xml:id="_NX"> </note>	фусноте

Будући да су многе публикације објављиване у периодици излазиле у неколико (узастопних) бројева, редовна пракса при приређивању часописа у периоду обухваћеном пројектом била је да се читаоцима кратким напоменама (некад испод наслова, некад на крају текста), у загради, наговјести да појединачне публикације имају наставак, или да се у конкретном броју завршавају. Неке од тих напомена (Слика 5.) биле су: (Свршиће се), (Мјесто свршетка наставак), (Продужиће се), (Наставиће се), (Наставак) и (Свршетак).



Слика 5. Напомене на крају текста публикација које су у периодици објављиване у више (узастопних) бројева

²¹ Број стране.²² У случају да се ради о периодично објављиваним публикацијама, X јесте број часописа, Y број стране.²³ У случају фуснота, X подразумева њихов редни број.

Како би ове напомене биле сачуване и у дигитализованом облику публикација, додјеливана им је етикета <trailer> <trailer>;

<trailer>(Свршиће се).</trailer>
<trailer>(Свршетак).</trailer>,
<trailer>(Мјесто свршетка наставак).</trailer>
<trailer>(Продужиће се).</trailer> <trailer>(Наставиће
се).</trailer> <trailer>(Наставак).</trailer>

Пошто путописи често обилују топонимима, ријечима и цитатима на страним језицима, те пропратним коментарима и објашњењима у виду фуснота, показали су се нарочито захтјевним током дигитализовања, што је подразумијевало уврштавање мноштва метаподатака у дигитални формат.

7. Публиковање припремљених дигитализованих облика и представљање корпуса научној јавности

У актуелном тренутку говоримо о Корпусу у настајању, имајући у виду да је завршна фаза у припреми за публиковање, у организацији Друштва за језичке ресурсе и технологије, у току. Иначе, Корпус је представљен научној јавности у току Међународне конференције *Јужнословенски језици у дигиталном окружењу (ЈуДиг)*, одржаној од 21–23. 11. 2024. године на Филолошком факултету Универзитета у Београду. Подразумијева се да ће након публиковања представљање научној јавности бити сврсисходније, будући да ће прикази укључити и конкретне примјере употребе или коришћења у научноистраживачке сврхе.

У будућим сродним подухватима планирано је проширивање Корпуса из истог периода новим жанровским компоненатама, књижевном критиком, кратким причама или информативним тектовима из доступних периодичних публикација, што би омогућило даље научноистраживачке активности базиране на ијекавском изговору, као равноправном, али запостављеном изговору српског језика. Такав истраживачки подухват имао би велики значај не само из угла лингвистичке оријентације, већ примарно екстралингвистичке, односно друштвено-историјске улоге дигитализованих публикација.

Захвалност

Овај рад настао је у оквиру пројекта под називом „Дигитализација српског књижевног наслеђа ијекавског изговора при Центру за дигиталну хуманистику Филозофског факултета Пале (прва фаза)“, који суфинансира Министарство за научнотехнолошки развој и високо образовање у Влади Републике Српске (број 19.032/961-119/23, од 29. 12. 2023. године).

Овом приликом захваљујемо колегама из Народне и универзитетске библиотеке Републике Српске, Народне библиотеке Србије, Библиотеке Филозофског факултета Пале, Универзитетске библиотеке „Светозар Марковић“, Народне библиотеке „Радислав Никчевић“ Јагодина, Библиотеке Сарајева²⁴, Националне и универзитетске библиотеке Босне и Херцеговине, Библиотеке Матице српске у Новом Саду и Народне библиотеке „Филип Вишњић“ из Бијељине.

*Прилог 1. Чланови пројектног тима који су учествовали у
анотирању и кориговању текста*

Члан тима	Број текстова	Број страна ²⁵
проф. др Јелена Марковић	4	69
проф. др Радославка Сударушић	6	115
проф. др Мирјана Лукић	8	92
доц. др Срђан Шућур	20	292
мср Душан Пејић	3	72
мср Борјан Митровић	3	69

²⁴ https://digital.bgs.ba/bosanke-mihajlo_milanovic_f/

²⁵ Како број словних знакова на појединачним странама није униформан, обим посла који је подразумевао дигитализовање публикација није нужно сразмјеран њиховом броју и обиму.

Извори

- [1] *Босанска вила: лист за забаву, поуку и књижевност*. Сарајево. Год. XVIII, св. 23/24 (31. децембра 1903). – Год. XIX, св. 13/14 (31. јула 1904).

Литература

- [1] Бабић, М. (2021). Статус српског језика и књижевности у Републици Српској. У М. Ковачевић (Ур.) *Зборник радова са Прве интеркатедарске србистичке конференције "Статус српског језика и књижевности у образовном систему"*, (63–80). Београд: Завод за унапређивање образовања и васпитања.
- [2] Ласица, Б. (2016). Српски књижевни часописи у БиХ у вријеме аустроугарске власти и њихов допринос у развоју просвјете, културе и националног идентитета. У М. Летић (Прир.), *Наука и евроинтеграције*. Књига 10. Том II (55–72). Пале: Филозофски факултет.
- [3] Ласица, Б. (2020). *Српска периодика у Босни и Херцеговини до 1918. године*. Пале: Српско просвјетно и културно друштво „Просвјета“.
- [4] Милановић, А. (2017). Теоријско-методолошки проблеми стварања корпуса славеносрпских текстова. У Р. Драгичевић и А. Милановић (Ур.) *Научни састанак слависта у Вукове дане*, 46/3, (137–145). Београд: Чигоја штампа.
- [5] Милосављевић, П. (2019). *О обнављању србистике*. Београд: Институт за политичке студије, Матица српска у Дубровнику; Грачаница: Логос.
- [6] Радоња, М. и Шућур, С. (2021). О Корпусу студената англистике (КорСАНг) и могућностима његове софтверске експлоатације. *Infotheca – Journal for Digital Humanities*, 21(1), 37–58.
- [7] Стојановић, Ј. (2010). Српски језик и државно-национални пројекти у 19. и 20. вијеку. У М. Ковачевић (Ур.). *Српски језик, књижевност, уметност - Зборник радова са међународног научног скупа одржаног на Филолошко-уметничком факултету у Крагујевцу: Језички систем и употреба језика*. Књига I, (11–30). Крагујевац: Филолошко-уметнички факултет Крагујевац, Скупштина града Крагујевца.
- [8] Шућур, С. (2022). *Трансфер у фразеолошкој компетенцији у писању на енглеском језику као страном код србофоних говорника*. (Докторска дисертација). Филозофски факултет Универзитета у Источном Сарајеву.

- [9] Bernard, L., Schöch, C. and Odebrecht, C. (2021). In Search of Comity: TEI for Distant Reading. *Journal of the Text Encoding Initiative*, 14, 1–21.

- [10] Hawkesworth, C. (2000). *Voices in the Shadows. Women and Verbal art in Serbia and Bosnia*. Budapest/New York: Central European University Press.
- [11] Jovanović v. Serbia, Application no. 41394/15, (2023).
<https://www.zastupnik.gov.rs/sr/база-праксе/пресудеодлуке-и-извршење/паун-јовановић-против-србије> (датум преузимања: 7. 3. 2025.)
- [12] Krstev, C. (2008). *Processing of Serbian – Automata, Texts and Electronic dictionaries*. Belgrade: Faculty of Philology, University of Belgrade.
- [13] Krstev, C. (2021). The Serbian Part of the ELTeC Collection Through the Magnifying Glass of Metadata. *Infotheca – Journal for Digital Humanities*, 21(2), 26–42.
- [14] Polomac, V., Kurešević, M., Bjelaković, I., Colić Jovanović, A., Petrović, S. (2023). Digitizing Cyrillic Manuscripts for the Historical Dictionary of the Serbian Language Using Handwritten Text Recognition Technology. *Slověne*, 12 (1), 295–316.
- [15] Schöch, C., Patras, R., Erjavec, T. and Santos, D. (2021). Creating the European Literary Text Collection (ELTeC): Challenges and Perspectives. *Modern Languages Open*, 1, 1–19.
- [16] Stanković, R., Krstev, C., Šandrih Todorović, B., Škorić, M. (2021). Annotation of the Serbian ELTeC Collection. *Infotheca – Journal for Digital Humanities*, 21(2), 43–59.
- [17] Stanković, R., Krstev, C. and Vitas, D. (2024). SrpELTeC: A Serbian Literary Corpus for Distant Reading. *Primerjalna književnost*, 47(2), 46–53.
- [18] Trtovac, A., Milnović, V., and Krstev, C. (2021). The Serbian Part of the ELTeC – from the Empty List to the 100 Novels Collection. *Infotheca – Journal for Digital Humanities*, 21(2), 7–25.

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

Srđan Šućur, Jelena Marković

Summary

As Meša Selimović argues in “The Fortress” that what is not written does not exist, so we might say that in the contemporary digital environment what is not digitised barely exists. Thus, the conversion of the analog into the digital, i.e., digital reformatting, has certainly become the cornerstone of collecting corpora all over the world in the growing field of Digital Humanities.

Within the general wave of digital reformatting, this article is aimed at presenting the Digitalisation of the Serbian Cultural Heritage of Ijekavian Pronunciation (1840–1920) with the Faculty of Philosophy Pale Digital Humanities Centre (stage one). The basic goal of this Project is collecting and digitising the Corpus of the Serbian Cultural Heritage of Ijekavian Pronunciation (1840–1920). The Corpus is in actual fact an addition to the SrpELTeC+ corpus, collected mainly within the COST action Distant Reading for European Literary History (CA16204). The SrpELTeC+ corpus, due to objective reasons, is predominantly Ekavian, consisting of more than a hundred novels in Serbian representing the cultural heritage of the targeted period (1840-1920) within the total collection of numerous European languages and cultures.

Bearing in mind that Ijekavian pronunciation is equal to Ekavian pronunciation of the Serbian language, we state that Ijekavian is underrepresented in the currently available Serbian corpora, and is consequently unavailable for both a wider readership and potential corpus research. Having noticed the necessity to fill in the Ijekavian digital corpora vacancy, we aim to initiate the collecting of complex corpora, starting with the current Project and the Corpus under construction, of the representative literary works written in the Ijekavian pronunciation of the Serbian language of the 19th and the beginning of the 20th century.

Initially, it was proposed that the Corpus of the Serbian Cultural Heritage of Ijekavian Pronunciation (1840–1920) with the Faculty of Philosophy Digital Humanities Centre should consist of 30 publications in the first stage. Owing to subsequent contextual factors, aimed at achieving the acceptable level of representativeness, the total number of publications

has reached 46, with 39 stories, 5 travelogues and 2 novels. During the process, several challenges have been met; the uneven availability of two-decade period publications, the scarcity of women writers, the text selection criteria, and selected text digitising. The Corpus now includes both canonical and non-canonical publications, e.g., stories by Petar Kočić, Svetozar Ćorović, Vaso Kandić, Marko Popović, Milan Trifunović, Joanikije Pamučina, among others; travelogues by Konstantin Hadžiristić, Jovo Besarović, Sava Kosanović, Risto Besarović, Sava Pješčić and Marko Car; and novels by Svetozar Ćorović and Radovan Perović Tunguz.

The travelogues were the most demanding publications in the digitisation stage, owing to numerous instances of metadata to be embedded into the digitised format (place names, foreign quotations, footnotes, chapters), solved by the available TEI Guidelines. Some stories were originally published in smaller, sequential instalments in periodicals, which also required thoughtful consideration. A metadata database, containing all the relevant information about the digitised titles has been created, since its existence is one of the preconditions that make corpora usable for scientific purposes.

The Corpus is purposefully labelled “stage one”, since we plan to enlarge it by future additions and expansions, using other available written genres, e.g., literary criticism and short stories. Such a collection of genre-based corpora would certainly enhance the visibility and availability of both the Serbian Cultural Heritage of the Period and Ijekavian Pronunciation of the Serbian language, which is a necessary step towards a complete chronological corpus-based description of Serbian culture, history and language.

Keywords: SrpELTeC+, digitalisation, Serbian literary heritage, Ijekavian Pronunciation

Дигитализација грађе Одбора за ономастику САНУ – значај, циљеви и први кораци*

Научни рад

DOI: 10.18485/judig.2025.1.ch13

Снежана Петровић¹,  0000-0003-1549-7294

Мирјана Петровић-Савић²,  0000-0001-6129-4741

Ана Шпановић³,  0000-0001-7368-9660

Ленка Бајчетић⁴,  0000-0002-2005-5426

Матија Нешовић⁵,  0009-0004-5822-6675

Јована Тодорић⁶  0009-0002-0706-6312

Апстракт

У раду су приказани значај, циљеви и први кораци у процесу дигитализације ономастичке грађе Одбора за ономастику САНУ, која садржи око 700 000 листића сакупљених са целокупног српског језичког простора у периоду од 1975. године до данас. Процес дигитализације укључује скенирање, документовање и моделирање различитих података записаних на листићима ономастичке грађе за потребе израде структуриране базе података и креирања корисничког интерфејса за приступ бази и уношење података. Приказан је поступак израде и имплементације интерфејса, а указано је и на проблеме и недоумице који су се јавили у овом процесу, као и на примењена

* Овај рад финансирао је Министарство науке, технолошког развоја и иновација Републике Србије према Уговору број 451-03-136/2025-03/ 200174, који је склопљен са Институтом за српски језик САНУ.

¹ Институт за српски језик САНУ, snezzanaa@gmail.com

² Институт за српски језик САНУ mirjana.petrovic@isj.sanu.ac.rs

³ Институт за српски језик САНУ tesicana@gmail.com

⁴ Иновациони центар Електротехничког факултета у Београду д.о.о., lenka.bajcetic@gmail.com

⁵ Институт за српски језик САНУ nesovic1998@gmail.com

⁶ Институт за српски језик САНУ jovanatodoric080@gmail.com

решења. Крајњи циљ дигитализације ономастичких записа јесте објављивање грађе у форми вишеструко претраживе платформе отвореног типа, доступне свима, како истраживачима тако и широј јавности. Платформа ће омогућити да се ономастички подаци организују, повежу и прикажу на више начина, укључујући различите врсте претрага и визуелизацију података у виду дигиталних мапа.

Кључне речи: српски језик, дигитализација, ономастика, израда интерфејса, визуелизација података.

1. Увод

У раду ће бити представљен пројекат дигитализације ономастичке грађе која припада Одбору за ономастику САНУ, као и први кораци у његовој реализацији. Оснивањем Одбора 1974. године створени су услови за систематска ономастичка истраживања на нашем језичком простору. Милан Будимир је био први председник Одбора, а након његове смрти (крајем 1975. године) па до 1981. године Одбором руководи Фанула Папазоглу, коју на том месту 1981. године наслеђује Павле Ивић (Стијовић С., 1997: 203). Рад на теренском бележењу српске ономастичке грађе иницирао је и осмислио академик Павле Ивић. У њима се први ангажовао и надаље имао главну улогу Светозар Стијовић сопственим сакупљачким радом и регрутацијом других сарадника (Петровић-Савић, 2023а: 604). Од 2004. на челу Одбора налази се академик Александар Лома, под чијим руководством су започети послови на дигитализацији ономастичке грађе.

Теренска истраживања започета су 1975. године на подручју Косова и Метохије. Прикупљени подаци из 652 насељена места, смештена у 25 општина, записани су на око 250 000 листића (Петровић-Савић, 2023а: 609).⁷ Истраживања су се истовремено одвијала и у другим деловима Србије, затим у Црној Гори, Хрватској и Босни и Херцеговини (Стијовић С., 1997: 204). Подаци сакупљени из тих области исписани су на око 500 000 листића, тако да данас картотека Одбора за ономастику САНУ укупно садржи око 700 000 листића.

На терену су прикупљане следеће врсте ономастичких података:

а) ојконими (имена насељених места);

⁷ Део грађе са Косова и Метохије објављивана је у часопису *Ономатолошки прилози*: Божовић, 1983; Букумирић, 1981, 1982, 1984, 1988, 1997, 2011; Пумпаловић, 2015, 2016; Стијовић Д., 1997, 1998, 2002; Стијовић С., 1979, 1981, 1982, 1984; Ћирић, 1989; Цветановић, 1988; Џоговић, 1983, 1985, 1986, 1987, 1996.

Мотила
Село се налази у ботачкој м-
вотиској равници откопајући велики
високих планина. Од Витина је уда-
љено око 5 км. Село је било старо, на-
селено староседелачким становниш-
твом. Једно је од највећих села у овом
крају.
Село има веома старо гробље,
са мноштвом старих надгробних

плоча. У центру се више не сахра-
њују. Зато гробље је новије датум-
на.

Црква св. Ђорђе је веома стара.
Скоро је реновирана и обновила
се зграда. Налази се на локалитету
Брест. Поред цркве је велика
камена и висока звоника. Свеке
славе су Ђорђица и Цариница.
Село чине Петроварадин

Слика 1а

Слика 1б

основну цркву. Лежа даље основну
цркву уле у суседном селу Врбовцу.
Из Мотила се од 1960. године одсе-
лило 150 етничких породица. Значај их,
има 118 двадесетих породица своје
цркви везује за Кучалово, а за Мот-
лаце, то једна за селом Трпаче и Вр-
бовцу, од осталих који да у цркви
лом баче.
Простором брзи знакови у цркви
је око 4,50.

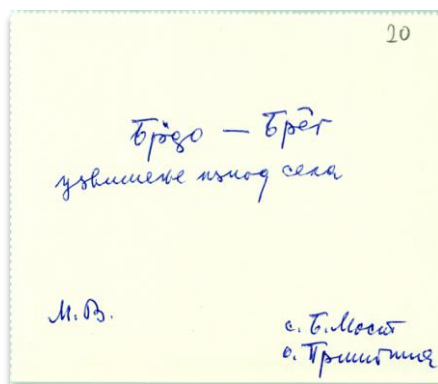
Слика 1в

б) етници (називи за становнике насељених места – мушки и женски) и ктетици (присвојни придеви изведени од имена места);

етници: Мотилци; Мотиланка;
ктетик: Мотиланка.
информатори: Петар Рачић (59),
Радослав Ђокић (68), Златко
Божик (65).
с. Мотила
о. Витина

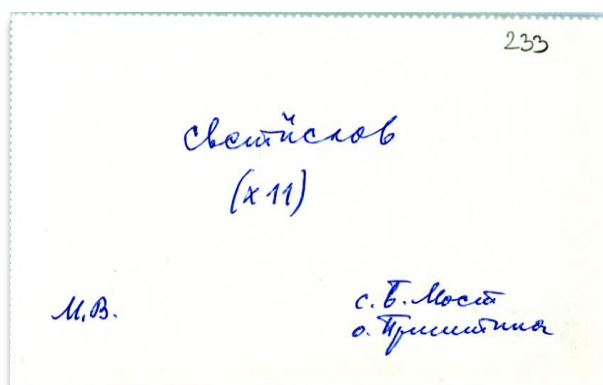
Слика 2

в) (микро)топоними;



Слика 3

г) антропоними (презимена, секундарна презимена, породични надимци, мушка лична имена, женска лична имена, мушки хипокористици, женски хипокористици, мушки лични надимци и женски лични надимци).



Слика 4

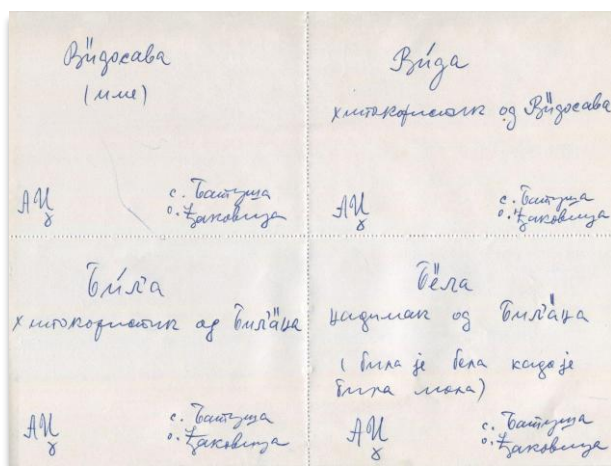
Битно је нагласити да нису у сваком пункту прикупљане све врсте података.⁸

⁸ Приликом бележења ономастичке грађе на терену у сваком посећеном пункту се, по правилу, бележе ојконими, етници, ктетици, микротопоними и патроними (уп. Стијовић & Гочанин, 2021: 294). Лична имена се не бележе у сваком селу. То је најчешће чињено у сваком петом селу једне области.

2. Процес дигитализације⁹

Први кораци у правцу дигитализације ономастичке грађе Одбора за ономастику САНУ начињени су пре петнаестак година, када је израђена база података и десктоп апликација за њихов унос.¹⁰

Наставак дигитализације састојао се из више фаза, од којих је прва скенирање целокупне грађе. Током 2020. године отпочеле су припреме листића за скенирање у Аудиовизуелном архиву и центру за дигитализацију САНУ, а приоритет је дат грађи са Косова и Метохије.¹¹ Будући да су ономастички подаци преко индига исписивани у блокове А4 формата са страницама подељеним на четири дела (Слика 5), листиће је најпре требало физички издвојити из ономастичких свезака, узбучити их и нумерисати. За ове послове у Одбору за ономастику САНУ било је ангажовано двоје хонорарних сарадника.¹²



Слика 5

⁹ Цео процес дигитализације обавља се уз финансијску и институционалну подршку САНУ, а у сарадњи са Аудиовизуелним архивом и центром за дигитализацију САНУ и Одсеком за дигитализацију Института за српски језик САНУ.

¹⁰ Сличан пројекат урађен је на Етимолошко-ономастичком одсеку Института за словеначки језик у Љубљани. Приређено је интернет издање рукописне збирке географских имена Макса Плетиершника (Pleteršnikova rokopisna zbirka zemljepisnih imen, <https://przzi.zrc-sazu.si/>), која представља изузетно богато сведочанство о словеначкој топонимији на прелазу из 19. у 20. век. Грађа из те збирке унета је у базу у виду скенираних листића, засад без прекуцавања ономастичких података у њој (Петровић-Савић, 2023б: 203).

¹¹ У међувремену је комплетан материјал припремљен за скенирање.

¹² Почетком 2022. године окончано је скенирање ономастичке грађе са терена Косова и Метохије. Очекује се да овај посао за грађу из остатка Србије и појединих делова бивше Југославије буде завршен током 2026. године.

Током скенирања су одређени подаци који се налазе на листићу – име села и општине којој оно припада, као и име и презиме сакупљача (на листићу исписано иницијалима) – везани за појединачне скенове и унети у Ексел табелу.

Стара база и интерфејс за унос података, иако у основи добро осмишљени, због ограничених техничких капацитета и немогућности превођења у веб-окружење, нису могли да испуне захтеве пројекта.¹³ Стога је следећа фаза дигитализације подразумевала израду нове базе и веб-апликације за унос ономастичких података са модерним корисничким интерфејсом који омогућава вишекорисничко окружење за унос, преглед и кориговање ономастичких података, као и података везаних за локације (географске координате, измењени називи места и општина и др.). Корисничка апликација је имплементирана помоћу програмског језика *Python* и библиотеке *Django*. У базу су иницијално унети метаподаци из већ поменутог Ексел документа.

Интерфејс је унапређен и пружа више могућности за унос и повезивање података, а самим тим и за њихову претрагу. На Слици 6 види се тренутни изглед корисничког интерфејса.

Слика 6

На левој страни екрана налази се скенирани листић. Изнад њега су поља „Редни број”, „Записивач” и „Локација”. У поље „Редни број” уноси се број са листића који је уписан приликом припреме

¹³ Прву верзију базе података и интерфејса израдио је Брана Томић.

грађе за скенирање.¹⁴ Поља „Записивач” и „Локација” попуњавају се избором из падајућег менија, начињеног на основу метаподатака из Ексел документа.

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 појам¹⁵ и коментар¹⁶. 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, имамо још два истог имена: „Потребна ревизија”. Овако названо поље смештено у горњем левом делу екрана омогућава да се назначи да ли је потребна исправка података преузетих из Ексел документа и везаних за скенирани листић.¹⁷ 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.

3. Проблеми и решења

Велики број проблема проистиче из недоследног и неуниформног бележења података на терену. Упутства за сакупљање грађе нису могла бити толико детаљна да обухвате све ситуације на које се наилазило у пракси, те су проблеми и недоумице решавани различито од случаја до случаја. Због тога смо се определили да грађу у базу уносимо прекуцавањем са листића уместо коришћења ОЦР (Optical character recognition) технологије. Пречишћавање података, ручна коректура и парсирање би у садашњим околностима одузели више времена него прекуцавање у појединачна поља, посебно имајући у виду да је максимална тачност података за лингвистичка истраживања услов без којег она не могу бити валидна.

Један од изазова представља и то што се један ономастички податак не везује увек само за један листић. Сходно томе број листића неће одговарати броју ономастичких података, већ се очекује већи број података него листића. У наредним примерима представимо неке од

¹⁴ То је једини податак који се налази на скенираном листићу, а није забележен на терену.

¹⁵ В. проблеме 2 и 3 у наставку.

¹⁶ У ово поље истраживачи могу унети текст који се не налази на листићу.

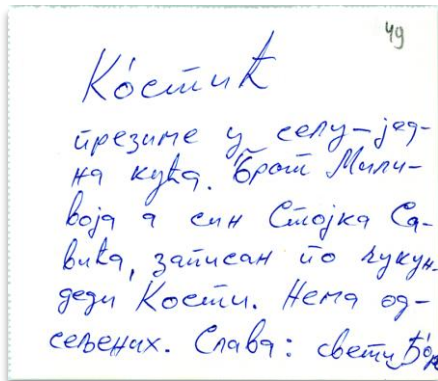
¹⁷ Може се, на пример, десити да је у Ексел документу унето погрешно име записивача теренске грађе, па је то неопходно изменити.

проблема с којима смо се сусрели, као и начине на које смо их разрешили.

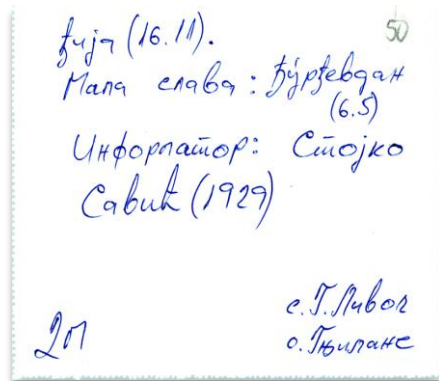
1. Један ономастички податак записан је на више листића (Слика 7а и 7б)

Грађа прикупљена на терену по правилу је уношена у ономастичке блокове тако што је на један листић уписиван један податак. Постоје, међутим, примери у којима су додатне информације уз један податак бележене на више листића (Петровић-Савић, 2023а: 612). То је углавном случај са информацијама о месту у којем се бележи ономастичка грађа (подаци о географском положају насеља у односу на његов административни центар, подаци о броју становника и његовом етничком саставу, о конфигурацији терена, о томе има ли место школу, цркву и слично). Понекад се податак код презимена исписује на више листића и укључује информације о броју породица које то презиме носе, њиховој крсној слави, одакле су досељени и тако даље.

У случајевима овога типа од више листића прави се један ономастички податак.



Слика 7а



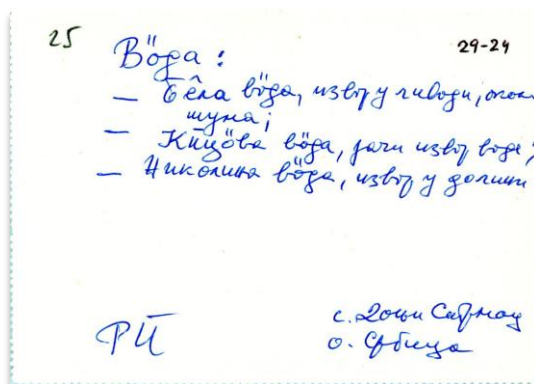
Слика 7б

2. На једном листићу налази се више потпуних ономастичких података (Слика 8)

У примеру на слици насловна реч *Вѡда* нема опис, што значи да на терену не постоје ни једночлани хидроним, нити топоним *Вѡда*, већ само двочлани називи синтагматског типа са лексемом *вѡда* (*Бѣла вѡда*, *Кѣцѡва вѡда*, *Николина вѡда*).¹⁸

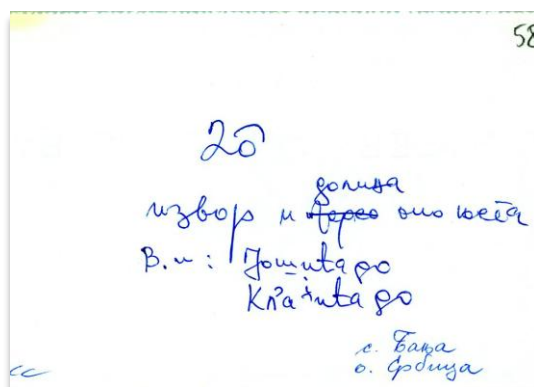
¹⁸ Исто важи за све насловне речи које немају опис, без обзира на то да ли су акцентоване или не.

У овом случају направљена су три ономастичка податка од једног листића: *Бела вџа*, *Кљџова вџа*, *Николина вџа*, а насловно *Вода* уписано је у поље „Кровни појам” уз сваки податак. Да је уз наслов одреднице стајао опис (нпр. *Вџа* „извор на одређеном локалитету”), од датог листића био би оформљен и додатни, четврти ономастички податак (уп. Проблем бр. 3).



Слика 8

3. На једном листићу налази се један комплетан ономастички податак и више упућеница (Слика 9)



Слика 9

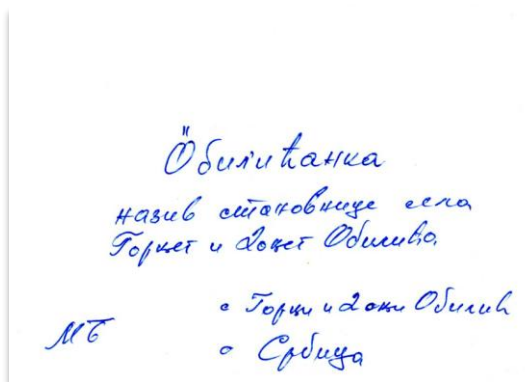
Код овог типа проблема насловна реч не представља кровни појам за вишечлане микротопониме чији је део, већ је самосталан ономастички податак са описом. Уз њега је записивач навео и друге

податке с напоменом „в. и” („види и”), чиме се указује на то да ће се они јавити и самостално, на посебном листићу.

У случају као на слици, микротопоним *Дѭ* обележен је као основни ономастички податак, док су називи *Јошића до* и *Кл'а'ића до* уписани у поље „Упућенице” и биће накнадно обрађени као засебни ономастички подаци када се њихови листићи буду појавили у апликацији.

4. Подаци из две локације записани су на једном листићу (Слика 10)

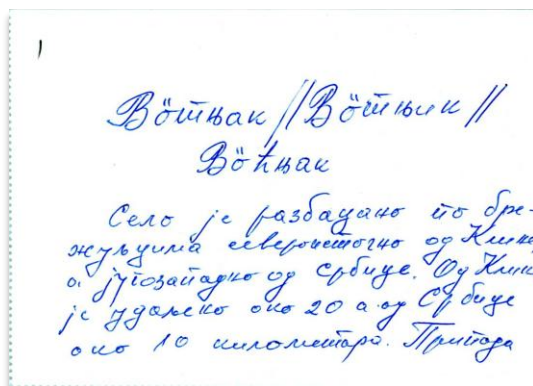
У случају када су на једном листићу записани ономастички подаци из два различита села, формирају се два одвојена ономастичка податка како би се податак потврђен у два локацијама (у примеру на слици *Ћбил'ићанка*) при претрази појавио у обе локације, чиме би се добио потпун резултат.



Слика 10

5. Насловна реч има више фонолошких варијанти (Слика 11)

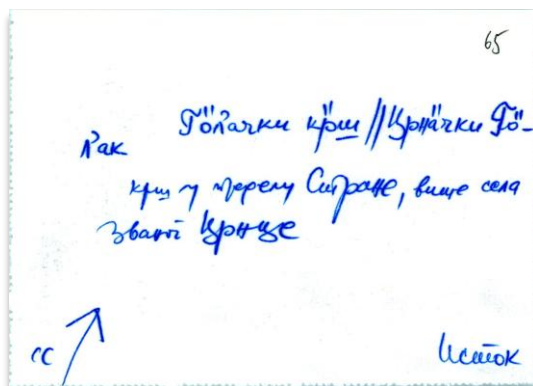
У случају када је на терену забележено више фонолошких варијанти насловне речи, у бази се формира један ономастички податак. Као примарни облик бира се онај којем је записивач дао предност при навођењу (на слици облик *Вѡтњак*), док се остале фонолошке варијанте (на слици *Вѡтњик* и *Вѡћњак*) смештају у поље „Алтернативни називи”. На тај начин ће корисницима бити омогућено да претрагом добију увид у све фонолошке варијанте насловне речи.



Слика 11

6. Локалитет има више различитих назива (Слика 12)

У случају да локалитет на терену има више од једног назива (на пример, микротопоним на слици забележен је као *Гол'ачки крш* и *Црн'ачки Гол'ак*) формирају се два одвојена ономастичка податка. Они се у апликацији обавезно везују и упућују један на други тако што се други податак уписује у поље „Упућенице”.



Слика 12

7. Недоследан начин бележења етника и ктетика (Слике 13а, 13б и 13в)

У инструкцијама за сакупљање грађе од записивача је тражено да бележе етнике и ктетике изведене од назива посећеног места, али не и од заселака који му припадају. Међутим, ово правило није увек поштовано, као што је случај у примерима на слици. Записивач је забележио два ктетика: *Кућичанин*, становник села Кућица, и

Доњострџмчанин, становник његовог засеока Доњи Стрмац. У оваквим случајевима, у апликацији ће се за сваки од ових назива формирати посебан ономастички податак.

1
КЋИЦА
Мешовито село у којем
има више српских него шип-
тарских кућа. Према броју
становника тај однос је у
корист Шиптара. Срби живе у
засеоку Доњи Стрмац (14 кућа),
али у том засеоку има и шип-
тара. У другом делу села, Кућ-
ци, живе само Шиптари. У селу
постоје три црквешта: једно
у српском делу (Црквене) код

Слика 13а

2 КУЋА Илија а два у шиптарском
делу на локалиту Здравс и дру-
го на локалитету Хасан-Мала.
Деца у школу одлазе у Горњи
Клину (само једно дете).
Према речима допашина, који
говори оба језика једнако и
чија је боја вокала шиптарско,
село се налази на тремеђи
Колашина, Метохије и Дрени-
це, али се води као крај Дре-
нице.

Слика 13б

3 ШИПТАРИ ЗАПИСАНЕ ГОЛО-
НИМЕ ИЗГОВАРАЈУ У ИСТОМ ОБ-
ЛИКУ КАО И СРБИ, С ЊИХОВИМ
ЈЕЗИЧКИМ ОСОБИНАМА.
ЕТНИЦИ:
КУЋИЧАНИН, КУЋИЧАЊКА
ДОЊОСТРМЧАНИН, ДОЊОСТРМЧАЊКА
ХТЕГЦИ:
КУЋИЧКИ
ДОЊОСТРМАЧКИ

Слика 13в

4. Закључак

У раду су приказани први кораци у дигитализацији ономастичке грађе и неки од проблема с којима смо се суочили током рада на њој. Наредне фазе дигитализације обухватиће даље унапређење апликације за унос података и уношење у базу свих података са Косова и Метохије, а затим и скенирање, обраду и унос преостале ономастичке грађе. Један од главних резултата пројекта биће визуелизација података на мапи у веб-апликацији, који ће претходно бити геолоцирани. Апликација ће имати кориснички

интерфејс који ће омогућити да обрађени подаци буду представљени на прегледан начин, као и ефикасну претрагу ономастичких података и листића. Рад на неким од будућих корака одвијаће се истовремено.

Оставља се могућност накнадног увођења ОЦР-а за рашчитавање података са листића, чему би морала претходити озбиљна анализа која би показала да ли би ова технологија уопште допринела ефикасности, брзини и прецизности уноса ономастичких података (имајући у виду различите рукописе бројних сакупљача, недоследности и грешке које су последица људског фактора и сл.).

Крајњи циљ овог пројекта јесте стварање свеобухватног дигиталног ономастикона српских народних говора који ће корисницима омогућити да грађу претражују по више различитих параметара. С обзиром на то да ће овај ресурс бити јавно доступан, послужиће широј публици као поуздан извор информација о нашим географским и личним именима и истовремено представљати драгоцен материјал за научна истраживања стручњацима различитих профила.¹⁹

Литература

- [1] Божовић, М. (1983). Ономастика села Бановог Дола код Звечана. *Ономатолошки прилози, IV*, 453–467.
- [2] Букумирић, М. (1981). Ономастика дела Прекорупља омеђеног рекама Мирушом, Белим Дримом и Клином. *Ономатолошки прилози, II*, 339–404.
- [3] Букумирић, М. (1982). Ономастика Прекорупља (II део). *Ономатолошки прилози, III*, 303–418.
- [4] Букумирић, М. (1984). Из ономастике јужне Метохије. *Ономатолошки прилози, V*, 413–603.
- [5] Букумирић, М. (1988). Ономастика Сиринићке жупе. *Ономатолошки прилози, IX*, 253–470.
- [6] Букумирић, М. (1997). Ономастика јужнога Косова. *Ономатолошки прилози, XIII*, 339–495.
- [7] Букумирић, М. (2011). Из ономастике Липљанске околине. *Ономатолошки прилози, XXI*, 47–190.
- [8] Лома, А. (2013). *Топонимија Бањске хрисовуље*. Београд: Српска академија наука и уметности.

¹⁹ Ономастичка грађа забележена на терену вишеструко је корисна за дијахронијски и синхронијски усмерена лингвистичка истраживања, као и за историјске, етнографске, културолошке и друге студије. У својим истраживањима историјске топонимије Александар Лома (2013) и Митар Пешикан (1981; 1982; 1983; 1984; 1986) обилато су користили податке сакупљене за Одбор са подручја Косова и Метохије.

- [9] Петровић-Савић, М. (2023а). Косовско-метохијска грађа при Одбору за ономастику САНУ: преглед обрађених и преосталих пунктова, послови скенирања, дигиталног уноса и интернет презентације. У: Д. Војводић (ур.), *Заштита, очување и афирмација српског културног наслеђа на Косову и Метохији, II* (стр. 603–624). Београд: Српска академија наука и уметности.
- [10] Петровић-Савић, М. (2023б). Pleteršnikova rokopisna zbirka zemljepisnih imen. *Зборник Матице српске за филологију и лингвистику, LXVI/1*, 203–205.
- [11] Пешикан, М. (1981). Из историјске топонимије Подримља. *Ономатолошки прилози, II*, 1–92.
- [12] Пешикан, М. (1982). Зетско-хумско-рашка имена на почетку турскога доба. *Ономатолошки прилози, III*, 1–120.
- [13] Пешикан, М. (1983). Зетско-хумско-рашка имена на почетку турскога доба. *Ономатолошки прилози, IV*, 1–135.
- [14] Пешикан, М. (1984). Зетско-хумско-рашка имена на почетку турскога доба. *Ономатолошки прилози, V*, 1–182.
- [15] Пешикан, М. (1986). Стара имена из доњег Подримља. *Ономатолошки прилози, VII*, 1–119.
- [16] Пумпаловић, Р. (2015). Ономастика Бањске и Звечана. *Ономатолошки прилози, XXII*, 599–738.
- [17] Пумпаловић, Р. (2016). Ономастика једног дела Ибарског Колашина. *Ономатолошки прилози, XXIII*, 369–511.
- [18] Стијовић, Д. (1997). Ономастика Суве Реке и околних села. *Ономатолошки прилози, XIII*, 497–573.
- [19] Стијовић, Д. (1998). Ономастика дела северне Метохије. *Ономатолошки прилози, XIV*, 157–256.
- [20] Стијовић, Д. (2002). Ономастика шест села у околини Витине. *Ономатолошки прилози, XV*, 227–263.
- [21] Стијовић, С. (1979). Ономастика источног дела Метохијског (Пећког) Подгора. *Ономатолошки прилози, I*, 237–380.
- [22] Стијовић, С. (1981). Ономастика средишњег дела Метохијског (Пећког) Подгора. *Ономатолошки прилози, II*, 197–338.
- [23] Стијовић, С. (1982). Ономастика западног дела Метохијског (Пећког) Подгора. *Ономатолошки прилози, III*, 193–300.
- [24] Стијовић, С. (1984). Ономастика једног дела поречја Кујавче. *Ономатолошки прилози, V*, 313–411.

- [25] Стијовић, С. (1997). Развој српске ономастичке науке у последњој четвртини века. *Јужнословенски филолог*, *LIII*, 203–210.
- [26] Стијовић, Р., Гочанин, М. (2021). Савремена ономастичка истраживања Косова и Метохије. У: Р. Младеновић, П. Пипер (ур.), *Резултати досадашњих и правци будућих истраживања српских народних говора Косова и Метохије* (стр. 291–309). Београд – Косовска Митровица: Српска академија наука и уметности – Институт за српски језик САНУ – Филозофски факултет.
- [27] Ђирић, Љ. (1989). Ономастика Изморника. *Ономатолошки прилози*, *X*, 365–496.
- [28] Цветановић, В. (1988). Из ономастике Гњилана. *Ономатолошки прилози*, *IX*, 517–542.
- [29] Џоговић, А. (1983). Топонимија југословенског дела Проклетија: (I део). *Ономатолошки прилози*, *IV*, 225–344.
- [30] Џоговић, А. (1985). Ономастика Дечана и околних села. *Ономатолошки прилози*, *VI*, 439–508.
- [31] Џоговић, А. (1986). Ономастика североисточних огранака Проклетија. *Ономатолошки прилози*, *VII*, 247–436.
- [32] Џоговић, А. (1987). Ономастика Ђаковице и суседних села. *Ономатолошки прилози*, *VIII*, 221–304.
- [33] Џоговић, А. (1996). Ономастика Горе. *Ономатолошки прилози*, *XII*, 33–366.

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

Snežana Petrović, Mirjana Petrović Savić, Ana Španović, Lenka Bajčetić, Matija Nešović, Jovana Todorić

Summary

The paper presents the significance, objectives, and initial steps in the process of digitizing the onomastic material of the Committee for Onomastics of the Serbian Academy of Sciences and Arts (SASA). The material comprises approximately 700,000 paper slips with data collected from the Serbian linguistic area, dated from 1975 to the present. The digitization process includes scanning, documenting, and subsequently modelling various data recorded on the paper slips to create a structured database and develop a user-friendly interface for accessing and entering data. The paper also outlines the procedure for designing and implementing the interface, highlighting the challenges encountered in this process and the applied solutions. The ultimate goal of digitizing onomastic records is to publish the material in the form of a multi-searchable, open-access platform available to both researchers and the general public. This platform will enable the presentation of onomastic data in multiple ways, including various search options and data visualization in the form of digital maps.

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

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

Scientific paper

DOI: 10.18485/judig.2025.1.ch14

Ranka Stanković,  0000-0001-5123-6273

Jovana Rađenović,  0000-0002-2707-3870

Maja Ristić,  0000-0002-3539-018X

*Dragan Stankov*¹  0000-0001-9394-9247

Abstract

The paper presents an overview of the different formats and domains of both multilingual and monolingual resources for the task of Question Answering (QA), with a particular focus on Serbian. Within the TESLA (Text Embeddings – Serbian Language Applications) project, a dataset consisting of context, questions, and answers is being developed with sources taken from various domains. The structure for the QA dataset, composed of four smaller subsets, and the method for its development will be outlined in this paper. The development of the first subset was based on an adaptation of a subset of the Stanford set SQuAD. The second subset *TESLA-Sveznanje-QA* was generated from the Sveznanje Encyclopedia, using LLMs, while the third part *TESLA-domain-QA* is produced from textbooks, also using LLMs. The fourth subset *TESLA-Wikidata-QA* contains automatically generated contexts based on the content of the Wikidata knowledge base and LLMs. The research conclusion indicates the importance and potential of applying this dataset in various fields, including educational technologies, digital assistants, and information retrieval systems. The presented results contribute to the improvement of language technologies for Serbian, and we hope that they will encourage further research and development in this area.

¹ University of Belgrade, Faculty of Mining and Geology

E-mail: {ranka.stankovic|jovana.radjenovic|maja.ristic|dragan.stankov}@rgf.bg.ac.rs

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

1. Introduction

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

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

The construction of high-quality question-answer (QA) datasets is essential for the development and evaluation of machine learning models, particularly for natural language understanding (NLU) and retrieval-based tasks. Traditional QA dataset creation methods rely on manual curation, rule-based extraction, or semi-automated approaches using structured knowledge bases. However, recent advancements in Large Language Models (LLMs) provide new opportunities for generating diverse and contextually rich QA pairs from textual corpora. Chen et al. (2019) show that while existing n-gram based metrics (BLEU, ROUGE, METEOR, F1) are somewhat suitable for current QA datasets, they significantly limit the development of more complex, free-form QA tasks and do not always correlate well with human response, suggesting a need for new metrics such as adapted BERT-based approaches.

Question Answering (QA) systems enable users to retrieve information from various sources, including both structured and unstructured data in natural language. As a crucial component of Conversational AI, QA has led to the emergence of Conversational Question Answering (CQA), where a system needs to understand the context and engage in multi-turn interactions to address user queries effectively. While most existing research has focused on single-turn QA, multi-turn QA has gained increasing attention

due to the availability of large-scale datasets and advances in pre-trained language models. With a growing number of studies contributing to the field each year, there is a pressing need to organize and synthesize existing research to guide future work. Zaib et al. (2022) provided a comprehensive review of state-of-the-art CQA research, highlighting a clear shift from single-turn to multi-turn QA, which enhances Conversational AI in multiple ways.

Knowledge Graph Question Answering (KGQA) has seen growing interest, with numerous benchmarking datasets driving advancements in the field. While earlier benchmarks relied on Freebase and DBpedia, research has shifted toward Wikidata due to its superior structural validity. In response, a new multilingual, complex KGQA benchmark was introduced as part of the QALD-10 series, transitioning from DBpedia to Wikidata. This adaptation required addressing challenges such as data complexity, cross-language mapping, and property ranking. Usbeck et al. (2024) presented a case study, conducted as a conference challenge, providing insights into the benchmark's creation and its role in advancing KGQA research. Knowledge Base Question Generation (KBQG) involves generating questions from structured database information, typically represented as triples. Hun et al. (2022) work on KBQG using pre-training, a new (triple, question) dataset, and question-type classification, demonstrating improved performance in standard and zero-shot settings. The extended KBQG2 dataset enhances knowledge base coverage and increases output variability, enabling multiple question formulations for the same KB triple.

Korablinov et al. (2020) introduced RuBQ, the first Russian Knowledge Base Question Answering (KBQA) dataset, containing 1,500 Russian questions of varying complexity, their English translations, SPARQL queries to Wikidata, reference answers, and a Wikidata sample with Russian-labeled entities. The dataset was created from online quiz question-answer pairs, which were subject to automatic filtering, crowd-assisted entity linking, automated SPARQL query generation, and in-house verification. RuBQ is a freely available and valuable resource for researchers in Semantic Web, NLP, and Information Retrieval, particularly in multilingual question answering. The dataset generation pipeline proved to be efficient and can be adapted for other data annotation projects.

QA datasets can be broadly categorized based on their structure, data sources, and the nature of the questions they contain. Extractive QA datasets, such as SQuAD³ and Natural Questions⁴, require models to identify answers

² The code and dataset are publicly available at <https://gitlab.inria.fr/hankelvin/wikidataqa>

³ <https://rajpurkar.github.io/SQuAD-explorer/>

⁴ <https://ai.google.com/research/NaturalQuestions>

directly within a given passage. In contrast, abstractive QA datasets, like NarrativeQA⁵ and certain parts of the MS MARCO dataset⁶, demand generative responses that go beyond exact text spans, requiring deeper language comprehension and rephrasing capabilities. Another major category is KBQA, where datasets such as WebQuestions⁷ and MetaQA⁸ rely on structured knowledge bases like Wikidata to retrieve factual answers.

With the rise of conversational AI, Conversational QA (CQA) datasets have become increasingly significant. Datasets like CoQA⁹ and QuAC¹⁰ enable models to engage in multi-turn interactions, maintaining context across multiple exchanges. Meanwhile, open-domain QA datasets, such as TriviaQA¹¹ and HotpotQA¹², allow models to search for answers across large, unstructured text corpora, often requiring retrieval from multiple sources. Multi-hop QA datasets, including HotpotQA and ComplexWebQuestions¹³, add another layer of complexity by requiring reasoning across multiple documents to arrive at a final answer.

Beyond factual retrieval, some QA datasets focus on commonsense reasoning, such as CommonsenseQA¹⁴ and Social IQA¹⁵, which challenge models to infer implicit knowledge beyond what is explicitly stated in text. Similarly, multimodal QA datasets, such as VQA¹⁶ and TextVQA¹⁷, extend the scope of QA to include images and documents, requiring an understanding of both textual and visual elements. Domain-specific datasets, such as biomedical QA datasets like BioASQ¹⁸ and MedQA¹⁹, address specialized fields where expert-level reasoning is required. Additionally, code-related QA datasets, such as CoNaLa²⁰ and CodeSearchNet²¹, are designed for programming-related queries, assisting in automatic code generation and problem-solving.

As QA research progresses, the diversity of available datasets continues to expand, enabling models to tackle increasingly sophisticated

⁵ <https://github.com/google-deepmind/narrativeqa>

⁶ <https://microsoft.github.io/msmarco/>

⁷ <https://github.com/brmsn/dataset-factoid-webquestions>

⁸ <https://github.com/yuyuz/MetaQA?tab=readme-ov-file>

⁹ <https://stanfordnlp.github.io/coqa/>

¹⁰ <https://quac.ai/>

¹¹ <https://nlp.cs.washington.edu/triviaqa/>

¹² <https://hotpotqa.github.io/>

¹³ <https://www.tau-nlp.sites.tau.ac.il/compwebq>

¹⁴ <https://www.tau-nlp.sites.tau.ac.il/commonsenseqa>

¹⁵ <https://maartensap.com/social-iqa/>

¹⁶ <https://visualqa.org/>

¹⁷ <http://textvqa.org/dataset/>

¹⁸ <https://participants-area.bioasq.org/datasets/>

¹⁹ <https://github.com/jind11/MedQA>

²⁰ <https://conala-corpus.github.io/>

²¹ <https://paperswithcode.com/dataset/codesearchnet>

questions. From simple fact-based queries to complex multi-step reasoning and conversational interactions, the development of high-quality QA datasets remains essential for enhancing the capabilities of AI-driven question-answering systems across various domains. Hopkins et al. (2019) presented results of the SemEval 2019 task on math question answering, utilizing a dataset derived from Math SAT practice exams with logical form annotations for a subset, which reached an accuracy of 45%, thus significantly outperforming a random guessing baseline.

As part of the TESLA (Text Embeddings - Serbian Language Applications) project, we are working on the preparation of a set of data: context, questions, and answers, collected from different domains. In this paper, we will explain several approaches that we are using for QA dataset development in the TESLA project. The approach for the development of the first dataset, based on the Stanford SQuAD set (Rajpurkar et al., 2018), will be described in Section 2. The TESLA-Sveznanje-QA dataset, derived from encyclopedic knowledge, will be presented in Section 3.

Section 4 is dedicated to TESLA-domain-QA, a domain-specific QA with a dataset containing at least 5000 questions with answers and the relevant context, excerpted from a textbook published by the Faculty of Mining and Geology. The last dataset, TESLA-Wikidata-QA contains automatically generated contexts based on the content of the Wikidata knowledge base and Gemini 4.0. Its development is explained in Section 5, which outlines a method for transforming structured knowledge into a high-quality question-answering system.

2. TESLA-SQuAD-sr Dataset

In the process of constructing the first dataset, we turned to the Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al., 2018) as the basic reference. SQuAD consists of textual contexts extracted from Wikipedia articles, with questions and answers manually created by human annotators based on the given context. This dataset is widely used for training and evaluating machine learning models in question-answering tasks, providing a well-structured benchmark with high linguistic quality.

Following this approach, we analyzed the SQuAD-sr dataset (Cvetanović & Tadić, 2024), which was developed by applying a semi-automated translation approach to the original SQuAD dataset. This process involved translating the English content into Serbian while ensuring that the structure of questions, answers, and contextual passages was preserved. The use of semi-automated methods allowed for a faster adaptation of an existing QA dataset into Serbian, while also introducing potential translation

inconsistencies that required manual correction. Table 1 provides the average (avg), maximum (max), and minimum (min) character count for each element in the SQuAD v1.1 and SQuAD-sr datasets. The primary purpose of these statistics is to enable a thorough analysis of text length within both datasets. This, in turn, facilitates a direct comparison of textual characteristics between the English and Serbian versions, offering insights into potential linguistic and/or structural differences.

Table 1. Dataset metrics SQuAD v1.1 and SQuAD-sr (in characters)

	SQuAD v1.1			SQuAD-sr		
	avg	max	min	avg	max	min
context	736	3706	151	715	3520	151
question	60	270	1	57	1063	0
answer	20	239	1	20	570	1

The first part of our dataset, *TESLA-SQuAD-sr*, represents a refined subset of SQuAD-sr, with corrections to improve translation accuracy, linguistic coherence, and domain relevance. The selected subset comprises topics such as information technology, climate change, civil engineering, geology, and similar areas. These topics were chosen to ensure a diverse yet specialized knowledge base that can support QA system development in both general and domain-specific contexts.

The finalized subset will contain approximately 7,000 questions with corresponding answers, offering a structured and high-quality dataset suitable for further model training and evaluation. This refined resource aims to contribute to the development of Serbian-language QA systems by enhancing both linguistic precision and contextual depth in machine-learning applications.

Currently, the dataset contains 5,000 QA pairs across 36 different topics. The original *SQuAD* dataset contains 87,599 QA pairs related to 442 topics, and the *SQuAD-sr* dataset contains 87,175 QA pairs for the same number of topics. Topics from *SQuAD* that weren't of interest, such as Sino-Tibetan relations during the Ming dynasty, The Legend of Zelda: Twilight Princess, Southampton, Catalan language, Estonian language, etc., were not taken into consideration for *TESLA-SQuAD-sr*. The questions in our dataset are formulated to cover different types of queries: questions that require specific facts, questions with descriptive answers (which require explanations or descriptions), and procedural questions, that is, questions that require a series of instructions or steps as a response. For every context, no more than five questions are included. For every question, one answer

is given. Data is collected in a variety of ways and verified through a manual annotation process to ensure the accuracy and relevance of responses. Tables 2 and 3 illustrate the need for manual correction e.g., named entities are written with errors (Applea instead of Eplovog), the word order is incorrect, and there are some grammatical errors.

Table 2. An example of context correction in the SQuAD-sr dataset

SQuAD	SQuAD-sr	Corrected SQuAD-sr
Apple's Safari had its first beta release in January 2003; as of April 2011, it had a dominant share of Apple-based web browsing, accounting for just over 7% of the entire browser market.	Safari od Applea imao je svoju prvu beta verziju u januaru 2003. godine; od aprila 2011. godine, imao je dominantan udeo Apel-baziranog veb pregledanja, što je činilo nešto više od 7% cele tržište pregledača.	Prva beta verzija Eplovog Safari pregledača objavljena je u januaru 2003. godine; od aprila 2011. godine, imao je dominantan udeo Epl-baziranog veb pregledanja, što je činilo nešto više od 7% celog tržišta pregledača.

Table 3. An example of a question and answer correction in the SQuAD-sr dataset

SQuAD	SQuAD-sr	Corrected SQuAD-sr
When was the first beta release for Safari?	Kada je prvi beta izdanje za Safari?	Kada je objavljena prva beta verzija Safarija?
January 2003	Januaru 2003.	U januaru 2003.

Current activities are focused on utilizing *Gemini 2.0* and *ChatGPT4* for automated grammar correction before manual review, to reduce the workload required for human correction. This approach enhances efficiency by minimizing the need for extensive human intervention while ensuring high linguistic accuracy and coherence in the final output. The lack of manually annotated datasets in Serbian makes these types of datasets particularly important.

3. QA Dataset from an Encyclopedia Using LLMs

This section outlines a method for generating a QA dataset from a structured encyclopedia containing tagged headwords and key entities. The goal is to leverage LLMs to generate high-quality question-answer pairs, ensuring that the dataset is both linguistically diverse and contextually accurate. The approach involves preprocessing, contextual understanding, question generation, answer extraction, and quality assessment. It is a work in progress, where the data has mostly been prepared and preprocessed, while the question and answer generation is still in the testing phase.

3.1. Data Preparation and Preprocessing

The encyclopedia used in this study is “Sveznanje”. A part of it, about 20%, is available online on Wiki Izvornik²², but the rest needs to be prepared from the paper copy. To that end, scanning was organized in the University Library Svetozar Marković, but after OCR, manual correction was also required. The next step was manual annotation of headwords and key terms marked with boldface in the original text. The part from Wiki Izvornik has 10,162 entries, while the digitized part currently comprises about 6000 entries, with the remaining entries to be finished by the end of this year.

The encyclopedia entries were segmented into smaller units based on paragraphs and headwords to ensure contextual coherence. They were structured using XML tags for entries, headwords, and key entities, and were used as input for data extraction and preprocessing to standardize the format for downstream tasks. In addition to the pre-tagged entities, a NER model is applied to identify additional significant terms. Entity linking is performed using external knowledge bases such as Wikidata to establish entity relationships. Any redundant or irrelevant text is removed, and duplicate entity occurrences are merged to prevent inconsistencies in question generation. TESLA-Sveznanje-QA is being generated using prompt engineering, where the LLM is given an encyclopedic article as context and asked to generate a set of questions and answers.

3.2. Question Generation Using LLMs

To produce diverse and contextually relevant questions, we utilize an LLM-based approach, which involves fine-tuning or prompting an existing transformer model such as GPT-4, Gemini 2.0, but we also plan to test other models in the future, like T5 and FLAN-T5.

Carefully designed prompts are used to instruct the LLM to generate questions based on encyclopedia content. Various strategies are explored, including:

- extractive QA generation, where the model is prompted to generate fact-based questions whose answers are directly found in the text;
- paraphrased QA generation, where the model is encouraged to generate multiple phrasings of the same question to enhance dataset diversity;
- reasoning-based QA generation, where questions requiring inferencing or logical reasoning beyond direct sentence extraction are included to improve depth.

²² <https://sr.wikisource.org/wiki/Свезнање>

For domain-specific texts, we plan to train and use a domain-specific LLM, where fine-tuning is performed using an existing corpus of high-quality QA pairs before generating new samples from the encyclopaedia.

3.3. Answer Extraction and Verification

Once questions are generated, answer extraction is performed using both automated and human-in-the-loop methods:

- extractive answer selection, designed for fact-based questions, with the LLM task to extract precise spans of text from the encyclopaedia;
- generative answering, for cases where extractive answers are insufficient, and the LLM should generate summaries or inferred responses.

The answer consistency check follows to ensure the accuracy. A secondary LLM pass is applied to verify that the generated answer is supported by the source text. Additionally, human annotators validate samples of generated QA pairs.

The finalized QA dataset is structured in a standardized format, typically JSON or TSV, with fields including:

```
{
  "context": "Text excerpt from encyclopedia",
  "question": "Generated question",
  "answer": "Extracted or generated answer",
  "answer_type": "Extractive / Generative",
  "headword": "Tagged headword",
  "entities": ["List of linked entities"]
}
```

For example:

```
{
  "context": "PROTON, jezgro vodonikovog → atoma, nosi izvestan naboj pozitivnog el. i ima 1847 puta veću masu od elektrona; u poslednje vreme se smatra da je složen (→ neutron); sastavni je deo atomskog jezgra svake vrste atoma.",
  "question": "Koliko puta je masa protona veća od mase elektrona?",
  "answer": "1847 puta",
  "answer_type": "Extractive",
  "headword": "PROTON",
  "entities": ["Q2294, Q9121, Q2225, Q2348"]
}
```

Quality assessment involves both automated metrics and human evaluation. The fluency of generated questions is assessed using language modelling metrics (perplexity and coherence). Cross-referencing with external knowledge bases should ensure that generated answers remain factually accurate. A set of annotators also assesses the dataset for readability, correctness, and ambiguity.

The presented method is an efficient approach to generating a QA dataset from an encyclopaedia using LLMs. The proposed pipeline ensures contextual accuracy, question diversity, and answer reliability. Future research will focus on incorporating multi-turn QA generation, adversarial question generation for robustness testing, and multimodal QA datasets integrating images and structured data.

4. Textbooks

The third dataset that is being developed, *TESLA-domain-QA*, will primarily focus on environmental protection, computer science, mathematics, mining, geology, and energy. It will consist of at least 5,000 questions along with corresponding answers and contextual passages extracted from textbooks published by the Faculty of Mining and Geology.

This dataset is designed to ensure high domain specificity and reliability, as the source materials (textbooks) provide well-structured and authoritative information. The selection of these fields reflects their growing importance in contemporary research and industry, making the dataset valuable for training and evaluating domain-specific question-answering models.

Table 4 presents an example of *context* in Serbian with three question-answer pairs. Translation to English is not a part of the dataset, and it is presented here for readability.

The inclusion of well-defined contexts ensures that the dataset aligns with real-world educational materials, enabling the development of AI models capable of answering questions based on structured academic knowledge. Additionally, this dataset can contribute to educational applications, intelligent tutoring systems, and specialized question-answering models in Serbian.

Table 4. An example generated from textbook

Context	Question	Answer
Koreni rudarstva na tlu današnje Srbije nalaze se u praistorijskom periodu. Arheološka istraživanja potvrđuju da je jedan od naših najstarijih rudnika metala, za koji se danas zna, rudnik bakra Rudna glava kod Majdanpeka. Otkriveno je i istraženo više okana i kanala odakle potiče obiman fond nalaza: keramički sudovi, kameni batovi–obluci različitih veličina i namena, koštani alati i dr.	U kom periodu se nalaze koreni rudarstva, na tlu današnje Srbije?	U praistorijskom periodu.
	Koji je jedan od naših najstarijih rudnika metala?	Rudnik bakra Rudna glava kod Majdanpeka.
	Koji su neki od pronalazaka na području rudnika Rudna glava?	Keramički sudovi, kameni batovi obluci različitih veličina i namena, koštani alati i dr.
The roots of mining on the territory of present-day Serbia date back to prehistoric times. Archaeological research confirms that one of our oldest known metal mines is the copper mine at Rudna Glava, near Majdanpek. Several shafts and tunnels have been discovered and studied, yielding a significant collection of artifacts, including ceramic vessels, stone hammers of various sizes and functions, bone tools, and other items.	In which period do the roots of mining on the territory of present-day Serbia date back to?	Prehistoric times.
	What is one of our oldest metal mines?	The copper mine at Rudna Glava, near Majdanpek.
	What are some of the discoveries in the area of the Rudna Glava mine?	Ceramic vessels, stone hammers of various sizes and functions, bone tools, and other items.

5. Wikidata

The fourth dataset, *TESLA-Wikidata-QA*, is also currently under construction, but the core components are already designed. The core of the *TESLA-Wikidata-QA* is GeminiKnowledge-sr²³, a large-scale, structured QA dataset in Serbian, comprising over 127,000 question-answer pairs. The dataset spans over a broad spectrum of general knowledge and is uniquely organized through a taxonomy, a three-layer hierarchical structure of ‘Area’ (26), ‘Topic’ (317), and ‘Facet’ (6376) (e.g., ‘Ekologija i održivost’ (‘Ecology and sustainability’) → ‘Biodiverzitet’ (Biodiversity) → ‘Tipovi ekosistema’ (‘Types of ecosystems’)). Linking the hierarchical structure with Wikidata is followed by the dataset generation as explained in this

²³ <https://huggingface.co/datasets/jerteh/GeminiKnowledge-sr>

section and in (Stanković et al. 2025). This granular organization facilitates fine-grained exploration, enables targeted training of models on specific knowledge domains, and allows for nuanced evaluation of model performance across different levels of specificity. Created to address the significant lack of publicly available, high-quality Serbian language resources, GeminiKnowledge-sr is provided in JSONL format to encourage its widespread adoption and utilization within the research community.

TESLA-Wikidata-QA will consist of automatically generated contexts derived from the content of the Wikidata knowledge base, aligned with GeminiKnowledge-sr. Unlike the previous datasets, which rely on manually curated or translated texts, this dataset relies on structured knowledge from Wikidata to create a question-answering (QA) dataset in a fully automated manner. Such an approach produces a scalable, diverse, and up-to-date resource for training and evaluating QA models.

The process of generating a QA dataset from Wikidata involves multiple steps, including data extraction, context generation, question formulation, answer extraction, and quality assurance. Wikidata is a structured knowledge base that stores factual information in the form of triples:

$$(\text{subject}) - [\text{predicate}] \rightarrow (\text{object})$$

For example:

```
(Nikola Tesla) - [occupation] → (inventor)
(Nikola Tesla) - [occupation] → (physicist)
(Nikola Tesla) - [field of work] → (electrical engineering)
(Nikola Tesla) - [notable invention] → (alternating current)
(Nikola Tesla) - [place of birth] → (Smiljan, Austrian Empire)
(Nikola Tesla) - [known for] → (Induction motor)
```

To extract relevant data, we use SPARQL queries to retrieve entities, properties, and relations related to target domains. The selection of entities and relations is guided by predefined topics, mentioned in the previous section. An example of SPARQL query to retrieve information about scientists and their fields of expertise is:

```
SELECT ?scientist ?scientistLabel ?field ?fieldLabel
WHERE {
  ?scientist wdt:P106 wd:Q901. # Occupation: Scientist
  ?scientist wdt:P101 ?field. # Field of work
  SERVICE wikibase:label { bd:serviceParam wikibase:language "sr". }
}
```

This query retrieves scientists and their respective fields of work, which can be used to generate contextual passages. We use Natural Language Generation (NLG) techniques to transform structured triples into coherent text passages. This transformation can be performed using predefined templates, such as:

- "[Subject] is known for [predicate] [object]."
- "[Subject] has made contributions to [predicate] in the field of [object]."

Alternatively, LLMs (e.g., GPT, T5) can be fine-tuned to generate more natural and varied contexts from structured data, but we are also experimenting with Gemini 2.0 for this task to obtain a greater variety of sentences and a more natural dataset.

Another example of using Natural Language Generation (NLG), where we transform Wikidata triples into a readable paragraph, is:

Extracted Data: *(Nikola Tesla) – [Known for] → (Alternating current, Induction motor, Wireless energy transmission).*

Generated Context: *"Nikola Tesla was a Serbian-American inventor, electrical engineer, and futurist known for his pioneering work in alternating current (AC) power transmission, the development of the induction motor, and early experiments in wireless energy transmission. His innovations in electrical engineering laid the foundation for modern power grids and influenced numerous technological advancements."*

Once the contexts are generated, questions are formulated based on the extracted information. The first approach for this step is template-based question generation, where predefined question templates are used to create structured questions:

- Who discovered [object]? → (Who discovered radium?)
- What is [subject] known for? → (What is Marie Curie known for?)

Another approach is LLM-Based Question Generation, where a fine-tuned question-generation model (such as T5) is employed to generate diverse and natural-sounding questions. The model is trained on existing QA datasets to learn how to rephrase and generate contextually appropriate questions.

Example of input-output:

Input (context from Wikidata):

"Marie Curie discovered radium and polonium, contributing significantly to the field of radioactivity."

Output (generated questions):

- *What elements did Marie Curie discover?*
- *Who contributed to the field of radioactivity?*

Answers are extracted directly from structured data in Wikidata. For example, if the question is “*Who discovered radium?*”, the corresponding Wikidata entity (*Marie Curie*) is retrieved as the answer.

There are two main strategies: 1) the answer is directly retrieved from the context, ensuring alignment with the original data; 2) if a simple answer does not fully capture the information, a generative approach (e.g., LLMs) is used to produce a more descriptive response.

We will illustrate here different types of generated questions:

- Fact-based: What is Nikola Tesla known for?
- Paraphrased: Which key inventions are attributed to Nikola Tesla?
- Specific: Who developed the alternating current system?
- Causal: Why is Nikola Tesla considered the pioneer in electrical engineering?

Using LLMs, the system can generate multiple phrasings of the same question, increasing dataset diversity.

An example of extractive vs. generative answers is:

- Extractive: “Marie Curie.”
- Generative: “Marie Curie, a physicist and chemist, discovered radium along with polonium in her groundbreaking research on radioactivity.”

More examples:

Question	Extracted Answer	Answer Type
Po čemu je poznat Nikola Tesla?	Naizmenična struja, indukcioni motor, bežični prenos energije.	Extractive
What is Nikola Tesla known for?	Alternating current, induction motor, wireless energy transmission.	
Gde je rođen Nikola Tesla?	Smiljan.	Extractive
Where was Nikola Tesla born?	Smiljan.	
Kakav je uticaj Nikola Tesla imao na elektrotehniku?	Razvio je sistem naizmenične struje, koji je postao osnova savremene distribucije električne energije.	Generative
What impact did Nikola Tesla have on electrical engineering?	He developed the alternating current system, which became the foundation of modern electricity distribution.	

To ensure accuracy and coherence, multiple automated and manual validation techniques are applied:

- fact verification, where answers are cross-checked with Wikidata and other knowledge sources,
- linguistic quality check, where generated contexts and questions are reviewed using grammar correction models (e.g., Gemini 2.0) before manual revision,
- diversity enhancement, where redundant or highly similar questions/answers are removed to ensure dataset variety,
- human annotation (sample review), where a subset of generated QA pairs is manually reviewed for clarity, factual correctness, and naturalness.

The final dataset is stored in a structured format, such as JSON or TSV, with the following fields:

```
{  
  "context": "Marie Curie discovered radium and  
polonium, contributing to radioactivity research.",  
  "question": "What elements did Marie Curie discover?",  
  "answer": "Radium and Polonium",  
  "answer_type": "Extractive",  
  "source": "Wikidata"  
}
```

This format ensures compatibility with machine learning models and ease of use for training and evaluation tasks. The automatic generation of a QA dataset from Wikidata provides a scalable and up-to-date approach to dataset creation, leveraging structured knowledge to produce high-quality questions and answers. Future improvements will include enhancing paraphrasing techniques to create more natural questions and expanding the dataset to include reasoning questions using entity graphs. Previous examples can be represented as follows:

```
{  
  "context": "Nikola Tesla was a Serbian-American  
inventor, electrical engineer, and futurist known for  
his pioneering work in alternating current (AC) power  
transmission, the development of the induction motor,  
and early experiments in wireless energy transmission."  
  ,  
  "question": "What is Nikola Tesla known for?",  
  "answer": "Alternating current, induction motor,  
wireless energy transmission.",  
  "answer_type": "Extractive",  
  "source": "Wikidata"  
}
```

6. Conclusion and Future Work

The application of this dataset, comprising four subsets, can be of substantial importance in various fields, including educational technologies, digital assistants, and information retrieval systems. By providing structured question-answer pairs, this dataset can serve as a valuable resource for improving machine comprehension and automated question-answering systems in Serbian.

Despite advancements in LLMs, even the most sophisticated models still lag behind human-level performance, highlighting the need for further development in the field of Natural Language Processing (NLP). The ability of AI systems to understand, reason, and generate linguistically and contextually appropriate responses remains an open challenge, particularly for languages with limited annotated resources.

A major issue in Serbian NLP research is the lack of manually annotated datasets, which restricts the ability to fine-tune models effectively. In this context, the dataset developed in this study can play a crucial role by bridging this gap and enabling the training of more accurate and context-aware AI models. By providing a high-quality, domain-relevant QA dataset, this research supports the development of Serbian language technologies, fostering improvements in machine translation, conversational AI, and educational tools.

Furthermore, this dataset may serve as a foundation for future advancements in language modeling, knowledge graph integration, and reasoning-based QA systems. We anticipate that this contribution will encourage further research and innovation in the domain of Serbian NLP, leading to the creation of more sophisticated AI-driven applications that can enhance human-computer interaction and knowledge accessibility.

Future work will focus on expanding the developed dataset by incorporating multimodal content (text, images, structured data), improving contextual understanding, and developing multilingual question-answering frameworks that enhance AI's ability to operate in diverse linguistic environments. By continuously refining and expanding this dataset, we aim to contribute to the evolution of Serbian AI systems, making them more robust, intelligent, and applicable to real-world scenarios.

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

References

- [34] Rogers, A., Gardner, M., & Augenstein, I. (2023). QA dataset explosion: A taxonomy of NLP resources for question answering and reading comprehension. *ACM Computing Surveys*, 55(10), 1–45.
<https://arxiv.org/pdf/2107.12708>
- [35] Rajpurkar, P., Jia, R., & Liang, P. (2018). Know what you don't know: Unanswerable questions for SQuAD. *arXiv preprint arXiv:1806.03822*.
<https://rajpurkar.github.io/SQuAD-explorer/>
- [36] Cenić, A. B., & Stojković, S. (2023). A Serbian question answering dataset created by using the web scraping technique. In *2023 58th International Scientific Conference on Information, Communication and Energy Systems and Technologies (ICEST)* (pp. 147–150). IEEE.
<https://doi.org/10.1109/ICEST58410.2023.10187370>
- [37] Chen, A., Stanovsky, G., Singh, S., & Gardner, M. (2019, November). Evaluating question answering evaluation. In *Proceedings of the 2nd workshop on machine reading for question answering* (pp. 119-124).
- [38] Cvetanović, A., and Tadić, P. 2023. “Synthetic Dataset Creation and Fine-Tuning of Transformer Models for Question Answering in Serbian.” In *2023 31st Telecommunications Forum (TELFOR)*, 1–4. IEEE.
- [39] Cvetanović, A., & Tadić, P. (2024). Synthetic dataset creation and fine-tuning of transformer models for question answering in Serbian. *arXiv preprint arXiv:2404.08617*. <https://arxiv.org/html/2404.08617v1>,
<https://paperswithcode.com/paper/synthetic-dataset-creation-and-fine-tuning-of>
- [40] Han, K., Ferreira, T. C., & Gardent, C. (2022). Generating questions from Wikidata triples. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 277–290). European Language Resources Association. <https://aclanthology.org/2022.lrec-1.29/>
- [41] Hopkins, M., Le Bras, R., Petrescu-Prahova, C., Stanovsky, G., Hajishirzi, H., & Koncel-Kedziorski, R. (2019, June). SemEval-2019 task 10: Math question answering. In *Proceedings of the 13th International Workshop on Semantic Evaluation* (pp. 893-899).
- [42] Korablinov, V., & Braslavski, P. (2020). RuBQ: A Russian dataset for question answering over Wikidata. In: Pan, J.Z., et al (Eds.), *The Semantic Web – ISWC 2020: 19th International Semantic Web Conference, Athens, Greece, November 2–6, 2020, Proceedings, Part II* (Vol. 12507, pp. 97–110). Springer.
https://doi.org/10.1007/978-3-030-62466-8_7
- [43] Stanković, R., Janković, N., Radenović, J., Ikonić-Nešić, M. (2025) From LLM Generation to Knowledge Representation: Creating and Structuring the GeminiKnowledge--sr QA Dataset for Serbian, *GOBLIN 2025: 1st GOBLIN*

Workshop on Knowledge Graph Technologies, Leipzig, Germany,
<https://www.dbpedia.org/events/goblin25-workshop/> (submitted).

- [44] Usbeck, R., Yan, X., Perevalov, A., Jiang, L., Schulz, J., Kraft, A., Möller, C., et al. (2024). QALD-10–The 10th challenge on question answering over linked data: Shifting from DBpedia to Wikidata as a KG for KGQA. *Semantic Web*, 15(6), 2193–2207. <https://doi.org/10.3233/SW-233471>
- [45] Zaib, M., Zhang, W. E., Sheng, Q. Z., Mahmood, A., & Zhang, Y. (2022). Conversational question answering: A survey. *Knowledge and Information Systems*, 64(12), 3151–3195. <https://doi.org/10.1007/s10115-022-01744-y>

Kreiranje skupa za obučavanje modela za odgovaranje na pitanja na srpskom jeziku

Ranka Stanković, Jovana Rađenović, Maja Ristić, Dragan Stankov

Sažetak

Razvoj i primena veštačke inteligencije u jezičkim tehnologijama značajno su napredovali poslednjih godina, posebno u domenu zadatka odgovaranja na pitanja (Question Answering - QA). Dok su postojeći resursi za QA zadatke razvijeni za glavne svetske jezike, srpski jezik je relativno zanemaren u ovoj oblasti. Ovaj rad predstavlja inicijativu za kreiranje obimnog i raznovrsnog skupa podataka za obučavanje modela za odgovaranje na pitanja na srpskom jeziku, koji će doprineti unapređenju jezičkih tehnologija za srpski jezik.

Pored brojnih istraživanja o jezičkim modelima u poslednjih nekoliko godina, mnogo je urađeno i na referentnim skupovima podataka potrebnim za praćenje napretka modeliranja. Posebno je puno urađeno kada je reč o odgovaranju na pitanja i razumevanju pročitane mada, uglavnom kada je reč o velikim jezicima (Rogers et al. 2023). U radu se pruža pregled različitih formata i domena raspoloživih višejezičnih i jednojezičnih resursa, sa posebnim osvrtom na srpski jezik (Cenić & Stojković 2023; Cvetanović & Tadić 2024).

U okviru projekta TESLA (Text Embeddings - Serbian Language Applications) radi se na pripremi skupa podataka: kontekst, pitanja i odgovori, prikupljenih iz različitih domena. Skup će biti sačinjen od četiri manja podskupa. U cilju izrade prvog podskupa, TESLA-SQuAD-sr, podskup Stanfordovog skupa SQuAD (Rajpurkar et al. 2018), gde je odgovor segment teksta, prevodi se i prilagođava, birajući teme kao što su:

Nikola Tesla, klimatske promene, građevina, geologija, itd. Podskup trenutno ima 5600, ali će imati oko 7000 pitanja sa pratećim odgovorima. Drugi podskup, TESLA-Sveznanje-QA, se priprema na osnovu enciklopedije Sveznanje na čijoj retrodigitalizaciji se intenzivno radi. Treći podskup, TESLA-domain-QA, koji se priprema na osnovu udžbenika u izdanju Rudarsko-geološkog fakultet, biće vezan pre svega za zaštitu životne sredine, informatiku i energetiku i sadržaće najmanje 5000 pitanja sa odgovorima i datim kontekstom ekscerpiranim iz udžbenika. Četvrti podskup, TESLA-Wikidata-QA, uglavnom će sadržati automatski generisane kontekste na osnovu sadržaja baze znanja Wikidata i Gemini 4.0 jezičkog modela.

Pitanja su pažljivo formulisana kako bi pokrila različite tipove upita: pitanja koja zahtevaju konkretne činjenice, pitanja sa deskriptivnim odgovorom (koja traže objašnjenja ili opis), i proceduralna pitanja, odnosno pitanja koja kao odgovor zahtevaju niz uputstava ili koraka. Podaci se prikupljaju na različite načine i verifikuju kroz proces ručnog anotiranja kako bi se obezbedila tačnost i relevantnost odgovora. Nedostatak ručno anotiranih skupova podataka na srpskom jeziku čini doprinos ovog istraživanja posebno značajnim.

Zaključak rada ukazuje na značaj i potencijal primene ovog skupa podataka u različitim oblastima, uključujući obrazovne tehnologije, digitalne asistente, i sisteme za pretragu informacija. Predstavljeni rezultati doprinose unapređenju jezičkih tehnologija za srpski jezik, i nadamo se da će podstaći dalja istraživanja i razvoj u ovoj oblasti.

Ključne reči: veštačka inteligencija, obrada prirodnog jezika, jezički resursi, anotirani skupovi, ekstrakcija informacija, odgovaranje na pitanja

Challenges and Perspectives in Italian Clitic Tagging: A Case Study on the SerbItaCor3 Corpus

Scientific paper

DOI: 10.18485/judig.2025.1.ch15

Saša Moderc¹  0009-0007-5059-2007

Abstract

Italian clitics are functionally heterogeneous: some function as pronouns, others as various types of adverbial adjuncts, or as markers of passive and impersonal constructions. Clitics are also used pleonastically and are integrated into pro-complement verbs (e.g., *farcela* ‘to manage’). They occupy three positions relative to the verb and can form clusters of two or three clitics; in such cases, the entire cluster usually occupies one of the three aforementioned positions.

In the SerbItaCor3_it corpus, the tagging of Italian texts was performed using TreeTagger (by Achim Stein). During analysis, inaccuracies were observed in the tagging of the clitics *ci* and *si*, as well as inconsistencies in the processing of homographs. The clitic *si* was tagged either as a reflexive pronoun (PRO:refl) or as a personal pronoun (PRO:pers), while its passive and impersonal uses were not marked. These tagging inaccuracies compromise the reliability of the corpus’s statistical data and limit its usefulness for linguistic analysis and language teaching. However, thanks to the accompanying Serbian translations, in most cases it was possible to determine the exact function of the clitics. This enabled the proposal of improvements to the Italian tagger, contributing to more accurate tagging of clitics and homographs. The paper presents examples of incorrect tagging and translation-based solutions, on the basis of which suggestions for improved tagging of Italian texts can be formulated.

Key words: Italian, Serbian, SerbItaCor3_it, homographs, clitics, tagging, improving tagger performances

¹ University of Belgrade – Faculty of Philology, moderc.sasa@gmail.com

1. Introduction: Italian Clitics

Italian clitics pronominalize grammatical persons, sentence arguments, or constituents. These include *mi*, *ti*, *ci*, *vi*, *lo*, *la*, *li*, *le*, *gli*, *ne*, and *si*. They are unstressed and bound to the verb or verb phrase, appearing in proclitic, enclitic, or mesoclitic position. Clitics form a prosodic unit with the verb and may be written either attached to it or separated from it, depending on their position. In terms of syntax, Italian clitics can occur in sequences of two, and occasionally even three, most commonly when *si* is used in impersonal or passive constructions (on these uses, a detailed discussion is given in Bentley 2006). The addition of a third clitic other than *si* is possible, though rare. Each clitic must have a unique referent (which may consist of one or more lexical items); when no referent is present, the clitic functions pleonastically or as part of a phraseme.

With certain frequent verbs, clitics may become lexicalized and significantly modify the verb's meaning. Such verbs are referred to as *verbi procomplementari* in Italian linguistics (see Russi 2008 for a detailed discussion). The syntax and functions of clitics are addressed in all Italian grammars, with a more extensive treatment in Renzi (1988), Serianni (1989), Salvi & Vanelli (2004) and Moderc (2021a, 2021b); most common uses of Italian clitics are discussed in a Serbian-Italian contrastive perspective in Moderc (2015).

2. The Polyfunctionality of Clitics

In standard Italian, clitics perform multiple functions, summarized in the following table (translated into English from Moderc 2021a: 21):

Table 1 Multiple functions of Italian clitics

FUNCTION	M I	T I	C I	V I	L O	L A	L I	L E	GL I	N E	S I
DIRECT OBJECT	●	●	●	●	●	●	●	●			●
INDIRECT OBJECT	●	●	●	●				●	●		●
REFLEXIVE FUNC.	●	●	●	●							●
PARTITIVE FUNC.										●	
SPATIAL FUNC.			●							●	
SOCIATIVE FUNC.			●								
INSTRUMENTAL FUNC.			●								
POSSESSIVE FUNC.										●	
IMPERSONAL FUNC.			●								●
PASSIVE FUNC.											●
PROFORM					●						
IDIOMATIC FUNC.			●		●	●		●		●	●

In addition to the functions mentioned above, other clitic functions emerge under the influence of colloquial language and dialects, or as a result of shifts in communicative strategy during spontaneous oral production. Despite these variations, the linguistic competence of speakers facilitates the correct association of clitics with their referents. Resolving clitic referentiality is a key component of language acquisition, as it requires an integrated understanding of syntax, verb valency, frequent collocations, and world knowledge, especially when context does not provide direct information about the referent and it must instead be inferred from extra-linguistic cues. Italian grammars and their accompanying exercises typically address only the most frequent clitic functions and the most common combinations of two clitics. However, language teaching demands a more comprehensive approach, aimed at enabling learners to identify each clitic's function and to substitute it with the appropriate referent or sentence constituent, typically expressed lexically (primarily with nouns).

In this context, annotated language corpora can facilitate the acquisition of clitic functions, as they are subject to automatic tagging. A considerable body of research has been conducted in this area (Schmid et al. 2007; Tamburini 2000, 2009; Dell'Orletta 2009; Schmid 2013), and the resulting findings are largely satisfactory, although they still show inaccuracies in the specific cases discussed in this paper. The tagging results can serve as a tool for testing and, if necessary, correcting learners' hypotheses about clitic functions in texts from the corpus. On this basis, we analyzed the extent to which the bilingual Serbian–Italian parallel corpus SerbItaCor3_it2 is reliable and accurate in identifying and distinguishing the functions of Italian clitics.

3. SerbItaCor3_it Corpus and Homographs: an Instrument for Successful Disambiguation

We begin our analysis by examining how homograph pairs are tagged in the corpus mentioned above. In the following examples, Italian nouns and verbs share the same form. Nouns are preceded by a definite article (*la* or *lo*), while verbs are preceded by a clitic (*la* or *lo* as unstressed personal pronouns). Since articles and clitics are themselves homographs, this results in what we may call “double homography” or “homographic syntagms”. The homographic syntagms in the following list represent an illustrative sample.³ In the English translations, nouns appear first, followed by verbs:

² More information in Modere S.; Stanković R.; Tomašević A.; Škorić M. (2023).

³ In English translations an indefinite article was preferred instead of the equivalent determinative *the*. In Italian there are no neutral nouns, therefore in some cases we had to use the pronoun *it*. Where needed, lexemes are added in order to stress the meaning or the English verb.

la caccia (a hunt; he/she chases her)
la cava (a quarry; he/she takes something out)
la guida (a guide; he/she guides her)
la leva (a lever; he/she removes her)
la manovra (a maneuver; he/she maneuvers it [e.g. *a car*]/he manipulates her [fig.])
la mostra (an exhibition; he/she shows her/it)
la piega (a fold; he/she folds it)
la posta (a mail; he/she posts it)
la sega (a saw; he/she saws it)
la sposa (a bride; he marries her)
la sveglia (a clock; he/she wakes her)
la veste (a dress; he/she dresses her)
le serve (the maids; she needs something)
il/lo perdono (a forgiveness; they lose it; I forgive him)⁴
lo sbaglio (a mistake; I get it wrong)
lo sconto (a discount; I discount [e.g. *this product by 10%*])
lo sfondo (a background; I break through [e.g. *the police checkpoint*])
lo sporco (dirt; I make/get it dirty)
lo sposo (a bride; I marry him/I'm marrying him)

Although some homograph pairs are correctly tagged exclusively as nouns (e.g., *sega* ‘a saw’ or *posta* ‘mail’, since the corresponding verb forms are not present in the corpus), or as verbs (since the corresponding nouns are not used), in several cases the tagging of homographs proves inaccurate. In a number of instances, *lo* is tagged as an article rather than as a clitic (i.e., a personal pronoun), even when it precedes a verb, a context in which determiners cannot appear. For example, the string “la cava” appears 14 times in the corpus SerbItaCor3_it: four times as a noun (‘a quarry’), once as a verb (‘to take something out’), and nine times as part of the idiomatic expression *cavarsela* (from *cavare* + reflexive *si* + idiomatic *la*, meaning ‘to manage, to get by’). Yet, in all 14 cases, *cava* is tagged as a noun (NOM). Correspondingly, the word *la* that precedes *cava* is always marked as an article (DET:det), which is only correct in the four cases where *cava* is a noun. In the remaining instances, *la* functions as a clitic, specifically, an unstressed personal pronoun without a definite referent, causing the verb to adopt an idiomatic meaning detached from any specific feminine singular noun as direct object. Similarly, the string “la conta” occurs seven times in the corpus SerbItaCor3_it. In one case, it is incorrectly tagged as a noun when it is actually a verb, as in *Chi non la*

⁴ In this case, two verbs are used: *pErdere* ‘to lose’ and *perdonAre* ‘to forgive’.

conta giusta a noi? (‘Who's not being straight with who?’, from *contare*, colloquial for ‘to tell’ + idiomatic *la*). Here again, *conta* is preceded by *la*, but both words are misclassified, *conta* as a noun [NOM] and *la* as an article [DET:def]. In the same corpus (SerbItaCor_it) the Italian string “le serve” is recorded 59 times; *serve* is dominantly tagged (56 times) as a noun (‘the maids’) preceded by the article *le*, and three times only as a verb (‘she needs something’), preceded by the personal clitic *le* (in the dative case, ‘to her’). In reality, the string “le serve” in 37 occurrences contains a noun and in 19 a verb, so that a revision of the POS tagging would be needed in this case also.

From these examples, it can be concluded that TreeTagger lacks the necessary instructions to distinguish between homographic nouns and verb forms. To enhance its performance, particular attention should be devoted to homographs, and specific rules or guidelines should be developed and implemented to facilitate their disambiguation. A possible control mechanism for this task could be derived from the data of the bilingual corpus SerbItaCor3_it, allowing an following a technical enhancement of the corpus itself. Such an enhancement would involve establishing links between semantically equivalent nouns, verbs, adjectives, and adverbs in the Italian and Serbian texts. After these adaptations, the SerbItaCor3_it bilingual corpus could be exploited to achieve a more accurate tagging of homographs, using the translations as reference points for refining the Italian tagger. For example, if the noun *serva* corresponds to *maid* or a similar term in Serbian (‘služkinja’, or a synonym), the tagging is appropriate; otherwise, it is reasonable to assume that *serve* represents a verb form of *servire*, as confirmed by the use of a verb in the Serbian translation (for instance, ‘služiti’), which reflects the actual function of this word in the original text. Although the list of homographs discussed here is not exhaustive, it nevertheless underscores that the issue has not been adequately addressed in the current tagging system, reducing the quality of linguistic annotation and, ultimately, distorting statistical data related to word classes. We assume that in the future, the integration of taggers with bilingual corpora and Large Language Models will allow for more accurate tagging of Italian words, while parallel bilingual or multilingual corpora will serve as valuable resources for verifying tagging accuracy. That said, we realize that the interpretation of ambiguous cases, such as *La porta la porta dal falegname*⁵ (‘He/She carries the door to the carpenter’), will

⁵ Depending on the interpretation and – in spoken language – on the intonation, “porta” can be interpreted in the first case as a noun and in the second as a verb, or vice versa, in the first case as a verb and in the second as a noun. The structure of the Serbian language in this case does not allow two different focalizations (left dislocation and right dislocation).

likely continue to require human intervention. Finally, we briefly mention the related issue of grammatical congruence in tagging. Specifically, the tagger should be instructed that a clitic cannot precede a noun, just as a determiner cannot precede a verb, except in the case of substantivized verbs, as in *Il bere fa male* (“Drinking is harmful to health”). Better results are obtained, for instance, in the tagging of the Paisà corpus (<https://www.corpusitaliano.it/>), where lemmatization and part-of-speech (POS) annotation, along with the indication of syntactic dependencies, have been applied. The TreeTagger used for the SerbltaCor3_it corpus, by contrast, does not appear to include syntactic dependencies.

This observation is prompted by the word *perdono*, which can mean either ‘forgiveness’ (in which case it is a masculine noun) or a verb form: ‘they lose’, from *pErdere*, or ‘I forgive’, from *perdonAre* (Italian accented vowels are represented by capital letters). In the corpus, we searched for the string “la perdono/i”, and in the six examples found, the tagger marked *la* as an article and *perdono/i* as a noun. This contradicts standard grammar rules, which require agreement in gender and number between the article and the noun (the correct forms being *il perdono*, *i perdoni*). Since in all six examples *perdono/i* is actually used as a verb, *la* preceding it can only be interpreted as a clitic pronoun (e.g., ‘to forgive her’, ‘to lose it’). In each corresponding Serbian translation, a verb is used to denote the action of forgiving or losing, rather than a noun. With the development of linguistic tools and the aforementioned corpus enhancements, translations into other languages (in this case, into Serbian) could serve as a valuable auxiliary resource for achieving a more accurate tagging of Italian parts of speech (POS).

4. SerbltaCor3_it Corpus and the Tagging of Clitics *si* and *ci*.

A search for the clitic *si* in the proclitic position in the “Ammaniti” subcorpus (part of the SerbltaCor3_it corpus), specifically within the novel *Io e te* (*Me and You*), yielded 217 results out of a total of 23,133 words. The clitic *si* is tagged either as a personal pronoun (*PRO:pers*, in 100 cases) or as a reflexive pronoun (*PRO:refl*, in 117 cases). In a sample consisting of the first 20 examples from the list of 217 results, the tagging was incorrect in six cases. In four instances (below, examples 3, 8, 13, 14), the label *PRO:refl* should have been used instead of the generic *PRO:pers* label; in two instances, *si* is an impersonal clitic, yet the tagger lacks a specific tag for this function. As expected, the dominant use of *si* is reflexive, corresponding to the *PRO:refl* label. However, the tagger also employs *PRO:pers* for the same reflexive function. To improve accuracy, two new

labels should be introduced: *PRO:imp* for impersonal *si* and *PRO:pass* for passive *si*. Undeniably, distinguishing between these two functions is not always straightforward, but successfully doing so would be a significant contribution to language teaching, especially by enabling the targeted extraction of sentences containing impersonal *si* and passive *si* constructions from the corpus. In the sample of 20 examples (fitting a single screen on the corpus interface), in two cases (4, 15), *si* was incorrectly tagged; the correct label for each is added in parentheses:

1. Dalla rabbia avevo preso un pietrone e l'avevo scagliato contro un albero, mentre quel ritardato **siREFL** rotolava a terra dalle risate. ✓
2. Mia madre e mio padre non lo sopportavano perché dicevano che **siREFL** prendeva troppe confidenze. ✓
3. Alla fine ha mollato la scopa e **siPERS (=REFL)** è avviato verso la guardiola con il suo passo dondolante e l'ho visto sparire sulle scale che portavano al suo. ✗
4. ... e al prato all'inglese con le panchine di marmo dove non ci **siREFL (=IMP)** poteva sedere. ✗
5. Due lunghi neon scarichi **siREFL** sono accesi illuminando un corridoio stretto e senza finestre...✓
6. La porta **siREFL** è spalancata su una grande stanza rettangolare...✓
7. ... un fluido rosso mi saliva per le gambe, mi inondava lo stomaco e mi **siREFL** irradiava fino alla punta delle mani...✓
8. Ma qui ci **siPERS (=REFL)** mettono tutti quelli che hanno problemi? ✗
9. mi avrebbe trasmesso, come un corpo caldo che trasmette calore a un corpo freddo, i pensieri dei bambini che **siREFL** erano sdraiati prima di me. ✓
10. Un Lorenzo che **siREFL** vergognava a parlare con gli altri ma che voleva essere come gli altri. ✓
11. Ho scoperto di avere un serbatoio nello stomaco, e quando **siREFL** riempiva lo svuotavo attraverso i piedi...✓
12. ... penetrava nelle viscere del mondo e **siREFL** consumava nel fuoco eterno. ✓
13. ... manager americani e italiani facoltosi che **siPERS (=REFL)** potevano permettere la retta. ✗
14. Uno **siPERS (=REFL)** è arrampicato sopra un albero e ha appeso lo zaino di una ragazza su un ramo e quella gli tirava le pietre. ✗
15. Chi aveva deciso che quello era il modo giusto ? Non **siPERS (=IMP)** poteva vivere diversamente? ✗

16. Io ho il sé grandioso, - ho sussurrato, mentre tre bestioni che **siREFL** tenevano a braccetto mi spingevano via come fossi un birillo...✓
17. I predatori in quella scuola erano molto più evoluti e aggressivi e **siREFL** muovevano in branco. ✓
18. Mi sono messo le stesse cose che **siREFL** mettevano gli altri. ✓
19. Il solco che mi divideva dagli altri **siREFL** faceva più profondo. ✓
20. E sotto la giacca dura come un esoscheletro **siREFL** agitavano cento zampette da insetto. ✓

In [21] another example of incorrect tagging of *si* is given. The appropriate annotation is provided in parentheses:

21. Gli unici rumori che **siREFL (=PASS)** sentivano erano la pioggia che batteva contro la finestra. ✕

With regard to the clitic *si*, we tested if TINT (The Italian NLP Tool, <https://dh.fbk.eu/research/tint/>, which allows users to test its functionalities in demo mode) would produce more accurate tagging results. We entered the Italian sentence: *Si dice che si sia convertito e a casa sua adesso si adorino gli idoli sumeri* ('They say that he converted and, in his home, now Sumerian idols are worshipped'). In all three instances, the clitic *si* was tagged identically (*Clitic=Yes, Person=3, PronType=Prs*), despite the fact that each *si* has a different function: impersonal (*Si dice*), reflexive (*si sia convertito*), and passive (*si adorino*), respectively. This uniform tagging implies that the user must manually determine the specific function of *si* in each context. To address this limitation, we recommend introducing distinct labels for the different uses of *si*, namely *PRON:Imp* (impersonal), *PRON:Refl* (reflexive), and *PRON:Pass* (passive), and enhancing the linguistic instructions required for a more advanced identification of each of these three functions.

As for the clitic *ci*, it appears 86 times in the aforementioned "Ammaniti" subcorpus. For the purposes of this study, we analyzed the first 20 occurrences in the list; inaccurate tagging is marked with the symbol ✕. We argue that tagging should distinguish among various functions of the clitic *ci*: locative (LOC), reflexive (REFL), phrasal or idiomatic (FRAS), pronominal (PERS), and, possibly, sociative (SOC) and instrumental (INSTR) uses. It can be assumed that proper differentiation of *ci* functions would necessitate tagging instructions accounting for syntax, semantics,

and textual coherence, an undertaking that is undoubtedly demanding and complex. In the following examples, we provide the correct label (or the most plausible one, in cases where the function of *ci* is ambiguous) in parentheses:

22. Così mi **ciREFL (=INSTR)** lavo e ti ho addosso. (clitic *mi* is reflexive; *ci* has an instrumental function, since it refers to a bar of soap, mentioned in the previous sentence) ✖
23. **CiREFL (=LOC)** hai messo dentro il termometro? ✖
24. Fortuna c'era un camion della spazzatura che **ciREFL (=PERS)** rallentava. ✖
25. Non avevo calcolato che mia madre **ciPERS (=FRAS)** tenesse tanto ad accompagnarmi. ✖
26. Non **ciREFL (=LOC)** vado. ✖
27. **CiREFL (=FRAS)** hanno messo un sacco a prepararsi... ✖
28. Allora **ciREFL** sentiamo stasera così la ringrazio. ✓
29. Il Cercopiteco **ciREFL (=FRAS)** ha messo parecchio a sentirlo. ✖
30. ...e al prato all'inglese con le panchine di marmo dove non **ciLOC** si poteva sedere. ✓
31. Ma quanto **ciREFL (=LOC)** devo stare? ✖
32. Ma qui **ciPRO:demo[nstrative] (=REFL)** si mettono tutti quelli che hanno problemi? ✖
33. Non **ciREFL (=FRAS)** voleva molto a fregarlo. ✖
34. Questo **ciREFL (=PERS)** sta dicendo il professore? ✖
35. Mi spiegava che gli amici **ciREFL (=FRAS)** mettono un attimo a dimenticarsi di te... ✖
36. Se **ciREFL (=PERS)** parla mia madre, - ha risposto Alessia Roncato. ✖
37. Io a Cortina **ciREFL (=LOC)** andavo da quando ero nato. ✖
38. Vedi che non **ciREFL** dobbiamo preoccupare. ✓
39. **CiREFL (=FRAS)** pensavo un po' e rispondevo tranquillo: «Va bene vengo». ✖
40. Mamma, ho deciso di non andare a sciare perché nonna sta male e se muore quando io non **ciREFL (=FRAS/LOC)** sono? ✖
41. Quanta neve **ciPERS (=LOC)** poteva essere? ✖

As shown by the examples provided, the clitic *ci* is tagged inaccurately, even more so than *si*. It is particularly surprising that the label for the reflexive function (REFL) is applied to verb forms that do not refer to the first-person plural, as *ci* can only function as a reflexive pronoun in this combination (e.g., *Ci troviamo bene a Pisa* ‘We feel comfortable in Pisa-). However, in this clitic-verb combination (followed by an object), a locative interpretation is also possible (e.g., *Ci troviamo un bel ristorante* ‘We find a nice restaurant there’). The inherent interdependence of context and the potential arguments of verbs further complicate the interpretation of *ci* and, as a consequence, its automatic tagging. Consequently, the development of more precise instructions for taggers is necessitated, and ultimately, human supervision appears to be needed.

5. Conclusion: why and how to Improve Programs for Tagging Italian Clitics

Improving taggers with more precise instructions regarding parts of speech and their functions may not be a primary focus in contemporary corpus linguistics. However, there is undoubtedly room for improvement in existing taggers and for applying a more advanced, detailed tagger in the next revision and expansion of the SerbItaCior3_it corpus, assuming such a tagger becomes accessible at that time. Greater accuracy in tagging and processing homographs and multifunctional words like clitics would enable linguists to conduct more focused and precise research within the corpus, allowing them to locate relevant examples more efficiently and test their hypotheses more effectively. In the field of foreign language teaching, the application of an improved tagger would provide corpus users with an efficient tool for extracting examples of specific clitic functions, as well as of other linguistic phenomena. This would make the corpus a more reliable learning tool, allowing students to observe a wide range of linguistic phenomena in Italian. Additionally, it would help clarify the usage and functions of clitics, particularly given the tendency in language teaching to overlook the complexities of referentiality and clitic functions (except for the most frequent uses and combinations), focusing instead on comprehension and production.

On the other hand, despite the mentioned flaws, we deem that the fact the parallel texts we have compiled over the years were integrated into the SerbItaCor3_it corpus is a great achievement for Serbian and Italian studies. This accomplishment is largely due to the efforts of Prof. Dr. Ranka Stanković, her associates, and their extensive experience in computational linguistics. Our comments on the tagging of the Italian portion of the corpus are intended to highlight areas where the corpus could be further improved; this task can be achieved by identifying and handling homographs,

especially in cases where the same form is used for different parts of speech (such as nouns and verbs, as demonstrated in this paper). Translations into Serbian can be especially useful in many cases, as shown with the homograph *serve*, which can correspond to either the Serbian noun *služkinja* ‘maid’ or one of the verbs expressing necessity (e.g. *služiti* ‘to serve’). For clitics, the Serbian language can serve as a control parameter in cases when clitics in both languages are used similarly (such as unstressed personal pronouns and reflexive pronouns).

The multifunctionality of Italian clitics and their dependence on context through referential relationships complicate the development of tagging instructions that can achieve a more reliable degree of accuracy. The congruence of the clitic *ci* with the first-person plural is a common example where *ci* functions as a reflexive pronoun. However, for some verbs, even this criterion is not entirely reliable. For instance, the phrase *ci troviamo* may mean ‘we are located’, or ‘we gather [in a place]’, or ‘we find ourselves [there]’ – in the last case a direct object is necessary, and this instruction should also be embedded in the tagger. Similarly, the reflexive pronoun *si* (which has distinct forms for each person) could be tagged according to its agreement with the corresponding person of the verb. However, even in this case, the third-person *si* can be mis-tagged when taken out of context. For example, in the phrase *Si dice bravo*, the verb can be either reflexive (‘He says of himself that he is good’) or impersonal (‘One says bravo’). The mismatch between reflexive verbs in Italian and Serbian represents an unreliable criterion for extracting tagging parameters, so this aspect of the tagger improvement would likely need to rely solely on elements of the Italian language, possibly with human revision to fine-tune the parameters.

Finally, given the complexity of tagging issues, a possible aid could lie in the integration of AI into the tagging process. In our experience, AI has proved to be enough accurate in identifying the functions of the clitic *ci*, as illustrated in the following response, obtained after a query asking ChatGPT to analyze the functions of the four occurrences of *ci* in the sentence below. The interpretations provided by ChatGPT are linguistically correct and are given in brackets:

42. **Ci** (IMPERSONAL, IDIOMATIC) vuole tanta fatica per riuscirci (PREPOSITIONAL OBJECT), ma se **ci** (LOCATIVE) vai e **ci** (PREPOSITIONAL OBJECT) provi, avrai successo. (‘**It takes** a lot of effort to succeed [**in it**], but if you go for it [litt. “**go there**”] and give **it** a try, you’ll succeed.’)

Further integration of AI with the SerbItaCor3_it corpus and its Serbian translations could prove useful in the tagging process.

References

- [1] Bentley, D. (2006). *Split intransitivity in Italian*. Berlin-New York. Mouton de Gruyter.
- [2] Dell’Orletta, F. (2009). Ensemble system for part-of-speech tagging. In P. Basile, F. Cutugno, M. N. Malvina, V. Patti, R. Sprugnoli (Eds.), *Proceedings of EVALITA 2009 – Evaluation of NLP and speech tools for Italian* (pp. 1-7). Reggio Emilia, Italy: EVALITA.
- [3] Moderc, S., Stanković, R., Tomašević, A., Škorić, M. (2023). An Italian-Serbian sentence-aligned parallel literary corpus. *Review of the National Center for Digitization*, 43, 1-20.
- [4] Moderc, S. (2015). *Gramatika italijanskog jezika: Morfologija s elementima sintakse*. Beograd. Luna crescens.
- [5] Moderc, S. (2021a). *I clitici italiani: Usi, ambiguità, interpretazioni. Volume primo: Il sistema dei clitici*. Beograd. Filološki fakultet.
- [6] Moderc, S. (2021b). *I clitici italiani: Usi, ambiguità, interpretazioni. Volume secondo: I nessi di clitici*. Beograd. Filološki fakultet.
- [7] Renzi, L. (1988). Renzi, L., Salvi, G., Cardinaletti, A. (eds.). *Grande grammatica italiana di consultazione* (Vol. 1). Bologna. Il Mulino.
- [8] Russi, C. (2008). *Italian clitics: An empirical study*. Berlin-New York. Mouton de Gruyter.
- [9] Salvi, G. & Vanelli, L. (2004). *Nuova grammatica italiana*. Bologna. Il Mulino.
- [10] Schmid, H., Baroni, M., Zanchetta, E., Stein, A. (2007). Il sistema "TreeTagger arricchito" – The enriched TreeTagger system. In B. Magnini & A. Cappelli (Eds.), *EVALITA 2007: Evaluation of NLP tools for Italian* (pp. 22-23). Retrieved from <http://www.evalita.it/2007/proceedings> [accessed October 24, 2024]
- [11] Schmid, H. (2013). Probabilistic part-of-speech tagging using decision trees. D. B. Jones, H. Somers (Eds.). *New methods in language processing*, No. 5. London. Routledge.
- [12] Serianni, L. (1989). *Grammatica italiana: Italiano comune e lingua letteraria*. Torino. Utet.
- [13] Tamburini, F. (2000). Annotazione grammaticale e lemmatizzazione di corpora in italiano. In R. Rossini (Ed.), *Linguistica e informatica: Multimedialità, corpora e percorsi di apprendimento* (pp. 157-171). Roma. Bulzoni.
- [14] Tamburini, F. (2009). PoS-tagging Italian texts with CORISTagger. In *EVALITA 2009: Workshop on evaluation of NLP and speech tools for Italian* (Vol. 1, pp. 1-7). Reggio Emilia, Italy: Accademia University Press.

Izazovi i perspektive tagovanja italijanskih klitika. Studija slučaja na materijalu korpusa SerbItaCor3

Saša Moderc

Sažetak

Italijanski klitici predstavljaju grupu reči koju karakteriše homogeno sintaksičko ponašanje i raznovrsnost funkcija. Njihova referencijalna polivalentnost, s jedne strane, i nekompletne instrukcije u jezičkim tagerima s druge, doprinose nedovoljno preciznom ili čak i pogrešnom označavanju klitika u jezičkim korpusima. Pored klitika, i homografi čine klasu reči koju tageri ne označavaju dovoljno precizno. Na primer, leksema *perdono* može imati značenje imenice ('oproštaj'), ali predstavlja i oblik glagola *perdonare* ('oprašati': *io perdono* 'ja opraštam'), odnosno glagola *perdere* ('gubiti': *loro perdono* 'oni/one gube'). Program TreeTagger, korišćen za morfološko označavanje reči u dvojezičnom korpusu paralelnih tekstova SerbItaCor3, ne raspolaže dovoljno preciznim instrukcijama za dodeljivanje ispravne oznake homografima. U ovom radu iznosimo pretpostavku da se dvojezični korpus može iskoristiti za podizanje preciznosti u tagiranju homografa, imajući u vidu da je polisemija iz italijanskog teksta leksički razrešena u prevodu na srpski jezik, te se odgovarajuće instrukcije za tagere mogu dedukovati iz korpusa SerbItaCor3. Označavanje polifunkcionalnih klitika *si* i *ci* takođe predstavlja izazov jer pomenuti tager ne sadrži instrukcije za preciznu obradu klitika. Usled strukturnih razlika između italijanskog i srpskog jezika, dvojezični korpus može samo u ograničenoj meri da pruži preciznije instrukcije za prepoznavanje specifičnih funkcija klitika. Stoga je realnija pretpostavka da se za italijanski jezik razviju posebni moduli za postojeći tager, sa posebnim instrukcijama za označavanje klitika *ci* i *si*. S obzirom na sintaksičku i semantičku složenost njihove upotrebe, neophodno je razmotriti i primenu manuelne provere i korekcije ispravnosti tagova koji se dodeljuju kliticima *ci* i *si*. Precizno razrešavanje homografa i ispravno tagovanje klitika, uz integraciju tagera sa potencijalima VI značajno bi povećalo pouzdanost lingvističkih podataka kojim je opremljen korpus SerbItaCor3 i doprinelo bi njegovoj većoj upotrebljivosti u lingvističkim istraživanjima i u didaktici italijanskog jezika.

Ključne reči: italijanski jezik, srpski jezik, SerbItaCor3_it, homografi, klitici, tagovanje, unapređenje tagera

Korpusi za učenje srpskog jezika kao stranog u eri veštačke inteligencije

Naučni rad

DOI: 10.18485/judig.2025.1.ch16

Olja Perišić¹  0000-0002-1219-862X

Apstrakt

U radu ćemo razmatrati izazove sa kojima se nastava srpskog jezika kao stranog zasnovana na korpusima suočava u kontekstu sve ubrzanijeg razvoja veštačke inteligencije. Dok se u oblasti Data Driven Learning-a sve više preispituje budućnost korpusnih alata (Crosthwaite, Baisa 2023, Flowerdew 2024), srpski jezik ostaje u povoju, barem što se tiče praktične upotrebe korpusa u nastavi. Naime, uočava se značajan jaz između istraživačkog delovanja stručnjaka za informatiku, posebno onih okupljenih oko udruženja JeRTeh, koji prate svetske trendove i aktivno razvijaju korpusne alate, i predavača koji bi te alate mogli da koriste u nastavi stranih jezika. Pored neophodnog afiniteta za ovu vrstu nastave, jedan od uzroka tog raskoraka je početno ulaganje u savladavanje veštine pretrage korpusima, kao i usvajanje teorijskog i metodološkog pristupa neophodnog za rad u učionici. U isto vreme iskustvo je pokazalo da studenti srpskog kao stranog relativno brzo savladavaju veštinu korišćenja korpusa i da su uglavnom kreativni u samostalnim istraživanjima, te time stiču motivaciju i sklonost ka istraživačkom radu i učenju jezika (Perišić 2023). Cilj rada je da se identifikuju i analiziraju ključni izazovi sa kojima se nastava zasnovana na korpusima u eri veštačke inteligencije suočava, sa posebnim osvrtom na generativne jezičke modele, kao i da se razmotri mogućnost njihove integracije u didaktiku srpskog jezika kao stranog.

Ključne reči: korpusi, Data Driven Learning, srpski kao strani, JeRTeh, veštačka inteligencija, generativni modeli

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

1. Uvod

Tokom tri i po decenije svoga postojanja pristup DDL (Data Driven Learning)² koji podrazumeva korišćenje korpusnih alata na svim nivoima usvajanja stranih jezika³ prošao je kroz različite interdisciplinarne razvojne faze na različitim poljima jezičkih istraživanja. Ubrzani razvoj digitalnih tehnologija i dostupnost kompjutera i mobilne telefonije sve većem broju korisnika već od početka novog milenijuma naveli su istraživače da preispituju efikasnost korpusa u odnosu na web alate i mogućnost njihovog integrisanja u didaktiku stranih jezika (Boulton 2012, Gatto 2014, Perišić 2025).

Od 2022. g. kada je ChatGPT ušao na velika vrata u naše živote i napravio revoluciju u mnogim profesionalnim sferama, uključujući obrazovanje, naučnici koji se bave DDL pristupom našli su se pred novim izazovom i potrebom da ispituju efikasnost i svrsishodnost korišćenja korpusa u eri generativnih modela zasnovanih na veštačkoj inteligenciji (Crosthwaite, Baisa 2023, Flowerdew 2024).

U domenu nastave srpskog jezika kao stranog, primećuje se značajan raskorak između istraživanja stručnjaka iz oblasti informacionih tehnologija, naročito onih okupljenih oko udruženja JeRTeh⁴, koji aktivno prate međunarodne trendove i razvijaju korpusne alate, i predavača koji bi te alate mogli primeniti u radu sa učenicima/studentima. Pored neophodne sklonosti ka ovom pristupu, jedan od glavnih razloga tog raskoraka leži u početnom trudu potrebnom da se ovlada veštinom korpusnog pretraživanja i da se usvoje odgovarajuće teorijske i metodološke osnove koje bi omogućile njihovu uspešnu upotrebu u nastavi. Od prvog istraživanja koje se bavilo potencijalom korpusa u nastavi srpskog kao stranog jezika (Miličević, Samardžić 2010), prošla je gotovo jedna decenija do pojave novih istraživanja u oblasti hrvatskog (Posavec, 2018; Mikelić Preradović, Posavec, Unić, 2019) i srpskog jezika (Perišić Arsić, 2018), koja, nažalost, još uvek ne pokazuju značajniju tendenciju rasta⁵. S obzirom na sve prisutniju ulogu veštačke inteligencije, posebno generativnih modela koji tektonski pomeraju način na koji percipiramo didaktiku i usvajanje jezika, cilj ovog rada je da preispita potencijal generativnih modela u didaktici srpskog jezika kao stranog kroz

² Osnivač ovog pristupa i onaj kome dugujemo njegov naziv bio je Tim Johns koji je devedesetih godina prošlog veka izveo prve eksperimente u učionici uz pomoć štampanih konkordanci sa studentima engleskog jezika kao stranog na Univerzitetu u Birminghamu (Johns, 1990).

³ U početku je fokus primene bio na akademskom nivou sa studentima srednjeg/višeg nivoa jezika, ali se vremenom polje njegove primene proširilo na sve nivoe jezičkog obrazovanja, ne samo na akademskom nivou, već i u srednjim i osnovnim školama (Braun 2007, Boulton 2020, Crosthwaite 2020, Hirata 2025).

⁴ JeRTeh - Language Resources and Technologies Society, <https://jerteh.rs/>.

⁵ Autorka rada se na svojim doktorskim studijama bavila ovom temom, koja je ostala u fokusu njenih istraživanja.

njihovo poređenje sa korpusnom metodologijom čija uloga u eri veštačke inteligencije mora biti preispitana, ali ne zanemarena.

2. Korpusi u eri veštačke inteligencije

Ostavljajući po strani značajna etička pitanja i pitanja regulacije upotrebe generativnih modela na različitim nivoima formalnog obrazovanja, jednostavnost ove vrste pretraga i činjenica da omogućavaju instant odgovore, kao i informativni i korektivni *feedback*, mogu biti odlučujući faktori kojima se vodi prosečan korisnik–student u odabiru jezičkih alata za učenje stranih jezika (Stewart 2025). Iz tog razloga potrebno je jasno preispitati prednosti i nedostatke ovih instrumenata u cilju njihovog boljeg razumevanja i eventualne implementacije u nastavu koja bi za osnovu imala kritičko preispitivanje, ali i eventualnu integraciju AI generativnih modela u već postojeća *corpus-based* istraživanja.

Gotovo svi autori koji su poredili generativne modele i korpuse naglasili su prednost korpusa u pogledu dostupnosti metapodataka (npr. izvora tekstova koji ulaze u njihov sastav⁶) (Crosthwaite, Baisa 2023) i autentičnosti jezičkih primera (Lin, 2023). U tom kontekstu postavlja se pitanje da li su automatski generisani tekstovi autentični s obzirom na to da nije dovoljno jasno na kojim su izvorima obučeni dok se generisanje teksta vrši na osnovu statističkih podataka i „embedding-a“. Pored toga u slučaju tzv. halucinacija, tj. generisanja netačnih ili izmišljenih informacija, pitanje je da li će nematernji govornici biti u stanju da samostalno prepoznaju neautentične ili lingvistički neodgovarajuće i neprikladne sadržaje. Još jedno pitanje tiče se mogućnosti replikacije kao osnovnog načela naučnog istraživanja, s obzirom na to da generativni modeli na isto pitanje mogu pružiti različite ili različito formulisane odgovore. Ukoliko pak poredimo generativne modele sa korpusima sa metodološke tačke gledišta, kriptičnost prvih po pitanju izvora i autentičnosti podataka čini *hands-on*⁷ pristup problematičnim, te je utisak da je u njihovoj trenutnoj razvojnoj fazi posredovanje nastavnika osnovni preduslov uvođenja u nastavu. Uloga predavača ne svodi se samo na tehničke aspekte nastave (npr. formulisane promptova), već i na ukazivanje na prednosti, ali i na nedostatke ovih modela na način koji podstiče kritičko razmišljanje i promišljenu i svrsishodnu evaluaciju rezultata.

⁶ Autor Kizito Tekwa (2024: 103) precizira da se radi o sledećim izvorima: „ChatGPT, for instance, uses corpus data from books, social media, Wikipedia, news articles, human speech and audio recordings, academic research papers, websites, forums, and code repositories“

⁷ Pristup *hands-on* podrazumeva direktno korišćenje korpusa ili u ovom slučaju generativnih modela, dok se pod pristupom *hands-off* podrazumeva štampanje output informacija poput konkordanci i njihova analiza i interpretacija (Boulton 2012).

U slučaju, pak, korpusne nastave učenik ne dobija gotove odgovore na istraživačka pitanja, već induktivnom metodom otkriva pravila funkcionisanja jezika ili deduktivnom metodom proverava postojeća pravila, u oba slučaja uz pomoć nastavnika, da bi vremenom sve više težio samostalnosti i autonomnom usvajanju jezika.

Pojedini autori predlažu poređenje radova pisanih od strane polaznika i automatski generisanih tekstova u cilju pronalaženja učestalih leksičkih formi koje bi ukazale na razlike u pismenoj produkciji čovek vs. mašina. Na osnovu kvalitativne analize tekstova studenti bi tokom nastave razvijali kritičko razmišljanje i imali uvid o potencijalu sopstvenih istraživanja koja mogu biti potpomognuta generativnim modelima, ali ne potpuno zamenjena njima (Zhang, Crosthwaite 2025). U tom smislu Brezina (2025) ukazuje na rizik površnog korišćenja AI instrumenata, jer imajući u vidu da se komunikacija odvija u različitim okolnostima i sa različitim komunikativnim ciljevima, podaci sami za sebe nisu dovoljni. U tom smislu autor tvrdi da je ljudski faktor i dalje nezamenjiv kada se radi o validnoj interpretaciji podataka i njihovom prilagođavanju preciznim kulturnim kontekstima i ciljevima.

Oba pristupa o kojima je reč u ovom radu (*AI-based* i *corpus-based*) u širem smislu deo su didaktičkog pristupa poznatog kao „self-directed learnig - SDL“ (Zhang, Zou, 2024) koji podrazumeva korišćenje alata u odnosu na kognitivni potencijal različitih profila polaznika. Naime, razlike među polaznicima nisu samo na nivou poznavanja jezika, već i u ličnim interesovanjima, ali i didaktičkim metodama koje su usvojili u prethodnom obrazovanju i koje nose sa sobom u vidu obrazovnog nasleđa, bez obzira na njihovu eventualnu efikasnost. S obzirom na to da su tradicionalni alati i kurikulumi standardizovani, ova vrsta personalizovanog pristupa obično je teško izvodljiva u tradicionalnoj učionici, što može ostaviti posledice kao što su nezadovoljavajući rezultati, sporo napredovanje, frustracije polaznika.

Za razliku od korpusa, generativni modeli omogućavaju korektivni feedback, mogu simulirati konverzaciju sa maternjim govornikom, drugim rečima mogu pomoći u razvijanju sposobnosti pisanja, čitanja, slušanja i sporazumevanja. Da bi ova vrsta interakcije bila efikasna potrebno je da korisnici imaju zadovoljavajući nivo AI pismenosti, odnosno da znaju na koji način da dođu do što boljih rezultata i da kritički procene efikasnost ovih alata koje ne treba prihvatati bezrezervno i bespogovorno. Interaktivno učenje koje nude ovi četbotovi nadovezuje se na konstruktivistički pristup u učenju jezika, jer omogućava da se kognitivni stil i pristup učenju svake individue valorizuje (Harel, Papert 1990). Ova vrsta konstruktivističkog pristupa učenju stranih jezika karakteristika je i

DDL-a koji podrazumeva aktivnosti „largely ustructured and open-ended“ koja zahtevaju aktivno učešće u nastavnom procesu, koji moduliraju na osnovu sopstvenih kognitivnih sposobnosti, interesovanja i metoda učenja (Flowerdew 2025).

3. Korpusi vs. generativni modeli u srpskom kao stranom jeziku

Prethodna istraživanja sprovedena na Univerzitetu u Torinu sa italofonima koji usvajaju srpski jezik kao strani, pokazala su da rad sa korpusima omogućava da se već od početnog nivoa razvija osećaj za jezik, istraživačka radoznalost i motivacija za učenje jezika (Perišić 2023). Zahvaljujući savremenim platformama za pretraživanje korpusa (Sketch Engine i NoSketch Engine⁸) studenti mogu da se obučavaju na različitim nivoima morfosintakse i leksike, objedinjenih prema principu *lexical grammar* (Sinclair 2000), istovremenog posmatranja i analize forme i značenja (*structure and meaning*). Na taj način se fokus pomera sa tradicionalnog učenja isključivo gramatičkih pravila na leksiku u kontekstu. Posebno su važna i inspirativna bila istraživanja koje su studenti radili samostalno, uglavnom na polju polisemije (razdvajanja bliskih sinonima), leksičkih praznina, kao i usvajanja glagolskog vida na višim nivoima, i sa derivacione i sa semantičke tačke gledišta. S obzirom na policentričnost nekadašnjeg srpskohrvatskog jezika, studenti su pokazali veliko interesovanje za usvajanje njegovih varijanti kroz širok spektar lingvističkih i sociolingvističkih pitanja (Perišić 2023). Ova vrsta istraživanja omogućena je pretragama tipološki srodnih korpusa poput onih zasnovanih na dokumentima preuzetim sa interneta za srpsku, hrvatsku i bosansku varijantu: SrWaC, HrWaC, BsWaC⁹ (Baroni 2009, Kilgariff et al 2010a).

Test sa kontrolnom grupom studenata koji nikada nisu koristili jezičke korpusne pokazao je da web pretrage mogu da zamene neke vrste korpusnih upita i na taj način dokažu da „googleology“ nije uvek „bad science“ (Kilgariff 2007), posebno na srednjem i višem nivou učenja jezika (Perišić 2025). Na ovim nivoima značaj interneta dodatno se proširuje u eri veštačke inteligencije, jer je svest učenika o jezičkim

⁸ Sketch Engine je trenutno najveća platforma za pretraživanje korpusa koja sadrži korpusne preko sto različitih jezika i jezičkih varijanti i pri tom omogućava pretragu i kreiranje korpusa putem prethodne registracije. Ova platforma je korišćena za pretragu korpusa SrWaC 1.2. (Ljubešić, Klubička, 2014)

NoSketch Engine je *opensource* verzija SE-a preko koje se mogu pretraživati korpusi za srpski jezik među kojima *SerbltaCor3* (Moderc et al, 2023) koji je korišćen za potrebe ovoga rada, a koji sadrži paralelizovane **srpske** i italijanske književne tekstove.

⁹ S obzirom na različit obim ovih korpusa u kvantitativnim istraživanjima studenti se obučavaju da koriste relativne frekvencije, a ne apsolutne.

strukturama autentičnog jezika već razvijena, dok je, kako je isto istraživanje pokazalo, na početnim nivoima potrebno opreznije pristupiti ovoj vrsti pretraga.

Da bismo proverili potencijal generativnih modela i uporedili ga sa korpusnim istraživanjima poći ćemo od tri već sprovedena istraživanja na polju polisemije (*kuća-dom*), leksičkih praznina (*babine*) i pozitivnih i negativnih konotacija reči tzv. semantičke prozodije (reč *žena* u kontekstu).

Praksa u učionici sa italo fonima koji usvajaju srpski jezik pokazala je da postojanje dva srpska prevodna ekvivalenta italijanske lekseme *casa* (*kuća* i *dom*) dovodi do nedoumica u njihovoj upotrebi, pri čemu se često pravi paralela sa engleskim rečima *house* (građevina) i *home* (emotivni, lični prostor). Istraživanje na Univerzitetu u Torinu pokazalo je da su studenti koji su prošli obuku korišćenja korpusa u stanju da samostalno utvrde polje primene ove dve lekseme na osnovu kolokacija i tako dođu do zaključaka o njihovom značenju i adekvatnom prevodu na italijanski (Perišić 2020).

U nastojanju da ovo istraživanje proverimo preko modela ChaGPT možemo formulisati sledeći prompt:

Koji su najfrekventniji pridevi koji formiraju kolokacije reči „kuća“? Nakon toga ponovi rezultate za reč „dom“!

Radi lakšeg poređenja, rezultati istraživanja preko korpusa *SrWaC 1.2* i *SerbItaCor3_sr* biće predstavljeni u tabeli zajedno sa rezultatima koje smo dobili preko generativnog modela ChatGPT.

Na prvom mestu uočava se da generativni modeli nisu u stanju da ponude tačne frekvencije reči s obzirom na odsustvo metapodataka vezanih za broj i vrstu tekstova na osnovu kojih je kompjuter generisao rezultate. Iz tog razloga ne možemo biti sigurni da li se zaista radi o najfrekventnijim ili najtipičnijim kolokatima. Osim toga, postavljanje istog prompta u različitim momentima daje barem delimično različite rezultate. Što se tiče sadržaja rezultata, uočava se da oba korpusna istraživanja ukazuju na značajnu komponentu institucionalne upotrebe lekseme *dom* (studentski, predstavnički, planinarski, parohijski, starački, Ruski, popravni), koje nam generativni model nudi samo u jednom slučaju (*dom* za stare). I u slučaju lekseme *kuća* korpusna istraživanja nude mnogobrojne rezultate ovoga tipa (izdavačka, gradska, medijska, Bela, robna, osiguravajuća, sigurna), potpuno odsutnih kod ChatGPT pretrage.

Tab. 1. Najfrekventniji pridevi lekseme kuća i dom (korpusi vs. ChatGPT).

	SrWac 1.2. ¹⁰	SerbItaCor3_sr ¹¹	ChatGPT
kuća	izdavačka (12.75), gradska (6.87), medijska (6.37), Bela (6.01), robna (4.53), osiguravajuća (3.98), sigurna ¹² (3.94), velika (3.59), porodična (3.53), stara (2.97)	izdavačka (9.34) ¹³ , cela (7.56), javna (7.23), velika (5.61), seoska (5.28), stara (5.28), nova (4.47), robna (4.23), roditeljska (3.49), lepa (3.33)	velika, mala, porodična, stara, nova, lepa, seoska, napuštena, skromna, drvena
dom	studentski (3.27), predstavnički (3.04), nov (2.95), planinarski (2.13), parohijski (1.89), donji (1.55), starački (1.14), vaš (1.13), Ruski (1.03), sav (1.02)	planinarski (3.09), starački (2.93), roditeljski (1.54), nov (0.98), popravni (0.81), studentski (0.81), parohijski (0.65), star (0.57), sopstveni (0.49), pravi (0.49)	porodični, topli, miran, roditeljski, siguran, sopstveni, ljubavni, stari (u značenju „dom za stare“), zajednički, pravi, mirni

Sa stanovišta maternjeg govornika nakon prvog dela može se zaključiti da je korpusna pretraga preciznija i verodostojnija, ne samo po pitanju frekvencija, već i po pitanju razdvajanja značenja dve lekseme zahvaljujući kolokatima. Ova vrsta informacija neophodna je stranom polazniku koji je sklon pravljenju paralela sa sopstvenim maternjim jezikom ili drugim jezicima koje poznaje, čak i u slučajevima u kojima zbog jezičke asimetrije to nije moguće.

U radu u kojem smo se bavili jezičkim prazninama (Perišić Arsić 2020) jedna od analiziranih leksema bila je *babine*, izabrana s obzirom na to da se radi o kulturološkoj praznini, u ovom slučaju običaju, koji ne postoji u italijanskom jezičkom i kulturnom prostoru. Sledi pretraga preko modela ChatGPT.

Prompt: Kako objasniti reč "babine/babinje" italijanskom polazniku?

¹⁰ Pri konsultovanju rezultata korišćena je funkcija Word Sketch.

¹¹ U nedostatku funkcije Word Sketch na NoSketch Engin-u korišćen je sledeći upit CQL: [tag="ADJ"] [lemma="kuća/dom"]

¹² Veći broj rezultata odnosi se na instituciju pod ovim nazivom koja nudi siguran krov žrtvama nasilja.

¹³ S obzirom na različitu veličinu korpusa unesene su relativne frekvencije koje su indikativne, jer funkcija Word Sketch nije potpuno ekvivalentna zadatom upitu. Na ovaj način nismo se ograničili samo na jednu vrstu pretraga, već su prikazane obe mogućnosti koje mogu biti korisne u budućim istraživanjima. Up. Kilgariff et al (2010b).

Odgovor:

To je običajni naziv za period neposredno nakon porođaja, kada se majka oporavlja i prima posete, a gosti donose poklone za bebu i porođilju. Obično traje oko 40 dana nakon porođaja, i vezan je za verovanja, brigu o zdravlju žene i deteta, kao i izbegavanje fizičkog napora.

Rezultat je veoma precizan i sa stanovišta maternjeg govornika srpskog jezika u potpunosti zadovoljavajući tj. smatramo da bi strani student na osnovu ovog objašnjenja i mogućnosti njegovog prevoda na maternji jezik bio u stanju da u potpunosti shvati značenje ove lekseme.

Pretraga preko ChatGPT modela po pitanju definisanja jezičkih praznina pokazala se veoma uspešnom, posebno imajući u vidu mogućnost generisanja sadržaja na maternjem jeziku polaznika. Prethodno korpusno istraživanje pokazalo je da je broj izvora koji sadrže ovu reč bio nedovoljan. Samim tim ova vrsta pretrage bila je delimično problematična, jer je u nedostatku dovoljnog broja rezultata u korpusu bilo neophodno poznavanje jezika na nivou maternjeg govornika da bi se definisalo značenje na osnovu raspoloživih kolokata.

Bilo bi važno sprovesti kompleksnije istraživanje na temu kulturološki specifične leksike koje bi obuhvatilo mnogo veći broj primera, diferenciranih po preciznim kriterijumima (običaji, narodne igre i plesovi, predmeti i slično) u cilju potvrde ovog preliminarnog rezultata koji ni u kom slučaju ne može poslužiti kao potvrda efikasnosti AI modela, već samo kao podsticaj za dalja istraživanja.

Pretraga reči *žena* u srpskom korpusu tekstova preuzetih sa interneta pokazala je da se ova reč često nalazi u negativnom kontekstu, pretežno uvredljivog ili violentnog sadržaja, od načina na koji se definiše karakter žene do nasilja koje se nad njom vrši. Korpusi su takođe omogućili poređenje ove reči sa njenim muškim ekvivalentom što je potvrdilo da je kontekst reči *muškarac* pozitivnije prirode, tiče se profesionalnih aktivnosti, snage, borbenosti, ali i seksualnosti¹⁴ (Perišić 2023). Akcenat kod žena je na reproduktivnoj ulozi i fizičkom izgledu, ali i na emocijama i profesijama kojima se bavi. Žena kao subjekt rađa, nosi dete, voli, zna, radi, živi i umire, dok se kao objekat u rečenici vezuje za glagole uglavnom negativne konotacije kao što su ubiti, voleti, silovati, tući, udati, zaposliti, varati, mrzeti.

Sledeće tabele omogućiće poređenje rezultata korpusnih istraživanja i ChatGPT pretrage:

¹⁴ Radi se o funkciji Sketch Difference prisutnoj na platformi Sketch Engine koja omogućava poređenje dve lekseme tako što svaku označava različitom bojom (crvenom i zelenom), te se time i na vizuelnom planu može bolje uočiti koji su kolokati zajednički, a koji ekskluzivni za svaku od reči.

Koji su najfrekventniji pridevi koji formiraju kolokacije reči „žena“?

Tab.2. Najfrekventniji pridevi, kolokati reči žena (korpusi vs. ChatGPT)

	SrWac	SerbItaCor3_sr¹⁵	Chatgpt-based
žena	mlada (6.11), sve (4.9), mnoge (3.48), lepa (3.15), poslovna (2.56), stara (2.35), dve (1.82), uspešna (1.13), prava (1.07), zaposlena (0.88)	mlada (22.43), druga (18.77), lepa (9.83), stara (5.93), prva (4.63), sirota (3.98), jedna (3.90), udata (3.66), mrtva (3.17), dobra (3.17)	lepa, pametna, snažna, dobra, mlada, trudna, starija, udata, nezavisna, uspešna

Drugi prompt je bio: *Koji su najčešći glagoli sa reči žena kao objektom u rečenici?*

Tab.3. Najfrekventniji glagoli koji imaju kao objekat reč žena (korpusi vs. ChatGPT)

	SrWac	SerbItaCor3_sr¹⁶	Chatgpt-based
žena	imati (1.75), voleti (0.81), videti (0.79), ubiti (0.42), naći (0.35), ostaviti (0.3), pitati (0.25), upoznati (0.25), tražiti (0.24), dovesti (0.23)	imati (3.74), ostaviti (1.87), pogledati (0.98), ubiti (0.81), naći (0.81), posmatrati (0.65), tući (0.57), nemati (0.57), videti (0.57), ugledati (0.49)	voleti, videti, upoznati, nazvati, trebati, želeti, poštovati, brinuti, braniti, hvaliti

Na prvi pogled uočava se da rezultati obe pretrage preko modela ChatGPT imaju pozitivnu konotaciju. Na našu opasku „Primećujem da nema glagola sa negativnom konotacijom“ ChatGPT odgovara: „Mnoge publikacije i mediji (koji pune korpusa poput srWaC-a) izbegavaju direktne izraze nasilja nad ženama (npr. tući ženu, ubiti ženu, silovati ženu) — posebno u eksplicitnom obliku.“ Jasno je da se radi o halucinaciji, odnosno netačnoj informaciji, jer rezultati do kojih smo došli, ni na koji način nisu izmanipulisani, niti softver za analizu korpusa nudi mogućnost filtriranja nasilnih sadržaja.

U ovom slučaju veoma je važno uočiti značaj istraživanja zasnovanih na frekvencijama, bilo da se radi o kolokacijama koje

¹⁵ CQL [tag="ADJ"][lemma="žena"]

¹⁶ [tag="VERB"][word="ženu"]

razgraničavaju značenje bliskih sinonima, bilo o onima čiji je cilj donošenje zaključka o semantičkoj prozodiji određene reči tj. pozitivnom ili negativnom kontekstu. Takav kontekst ne utiče samo na semantiku same reči, već u slučaju imenice *žena* oslikava društveni položaj žene, te ove vrste pretraga mogu da pomognu u razotkrivanju stereotipa i stigmatizacije pojedinih društvenih grupa. Kada je reč o naučnim istraživanjima koja zahtevaju preciznost i potkrepljenost informacija, može se zaključiti da korpusi zadržavaju svoju vrednost i značaj čak i u eri veštačke inteligencije, bar što se tiče trenutnog stanja u ovoj oblasti.

4. Zaključak

Polazeći od obrazovnih potreba novih generacija studenata, koji aktivno integrišu savremene tehnologije u sve aspekte svoga života, uključujući tu i obrazovanje, u radu smo preispitali efikasnost generativnih modela u odnosu na korpusne, kao i mogućnost njihove integracije u nastavu srpskog kao stranog jezika.

Posli smo od pozitivnih rezultata korpusnih istraživanja u ovoj oblasti i činjenice da se metodologija *corpus-based* uspešno koristi i pokazuje zadovoljavajuće rezultate na svim nivoima učenja jezika, posebno na srednjem/višem kada je svest o značaju autentičnog jezika već formirana. Ovo kratko istraživanje ograničili smo preciznim jezičkim pitanjima, imajući u vidu prethodna korpusna istraživanja u nastavi srpskog kao stranog na polju polisemije (*kuća-dom*), leksičkih praznina (*babine*) i pozitivnih i negativnih konotacija reči (*žena*).

Pokazalo se da je na polju kolokacija, u slučaju razgraničavanja značenja bliskih sinonima, korpusna pretraga preciznija od AI modela, jer omogućava kvantitativnu analizu na osnovu frekvencija, što nije moguće u slučaju generativnih modela zbog odsustva metapodataka. U isto vreme, kako je prethodno citirano istraživanje pokazalo, rezultati korpusne pretrage verodostojni su i mogu dovesti do novih saznanja, tj. pružiti uvid u nova značenja reči koja nisu dovoljno izdiferencirana u postojećim, posebno bilingvalnim leksikografskim izvorima za istraživanu jezičku kombinaciju.

U definisanju leksičkih praznina generativni model pokazao se potencijalno veoma funkcionalnim na svim nivoima usvajanja jezika s obzirom na to da se ovi sadržaji mogu prevesti na maternji jezik polaznika i time biti dostupni i studentima na nižim nivoima učenja jezika. Iako je i prethodno korpusno istraživanje dalo pozitivne rezultate u slučaju istražene lekseme treba imati u vidu da se radilo o pretrazi za koju je bilo neophodno solidno tehničko i metodološko predznanje, a koja bi eventualno mogla da

se proširi konsultovanjem novih paralelnih korpusa koje imamo na raspolaganju za srpski jezik (Perišić et al 2023, Moderc et al 2023). Da bi se proverili početni pozitivni rezultati generativnih modela u definisanju ove vrste leksike bilo bi neophodno sprovesti složenije istraživanje po pitanju kulturološki specifičnih pojmova izdiferenciranih prema strogim naučnim kriterijumima.

U slučaju reči s pozitivnim ili negativnim konotacijama pretraga pomoću generativnih modela otkriva brojne problematične aspekte. Naime, već je ukazano na to da ovi modeli podležu određenom obliku cenzure, kao u slučaju sadržaja nasilnog karaktera, pa stoga „izbegavaju“ reči s negativnim konotacijama. Zbog toga istraživanja iz oblasti analize diskursa, posebno ona koja se bave govorom mržnje ili ispitivanjem nasilnih i drugih negativno obojenih sadržaja trenutno ne daju validne rezultate.

U nastavi stranih jezika u Srbiji korpusi su nedovoljno zastupljeni čak i kada je reč o većim svetskim jezicima (Vitaz, Poletanović, 2019), što onemogućava sprovođenje složenijih kvantitativnih i kvalitativnih istraživanja koja bi ukazala na potencijalne prednosti i/ili nedostatke primene generativnih modela u glotodidaktici. Kao prvi korak neophodno bi bilo osposobiti što veći broj nastavnika, na različitim nivoima obrazovanja, za korišćenje korpusa u učionici i istovremeno pratiti rezultate dobijene putem generativnih modela, u okviru neke vrste integrativne metode zasnovane na principima DDL-a.

Predlaže se, dakle, usvajanje kritičkog mišljenja koje je u osnovi DDL pristupa i koje nije isključivo ograničeno na korpus, već kako je pokazano u prethodnim istraživanjima obuhvata i interakciju polaznika s web pretraživačima, ali i novim generativnim modelima. Ovakav integrativni okvir istraživanja može poslužiti kao osnova za definisanje didaktičkog modela koji bi se sprovodio sa studentima srpskog kao stranog jezika (i ne samo), sa ciljem razvijanja kritičkog razmišljanja na temu potencijala i efikasnosti generativnih modela na različitim nivoima usvajanja stranih jezika.

Literatura

- [1] Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2009). The WaCky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3), 209–226.
- [2] Boulton, A. (2012). What data for data-driven learning? *EuroCALL Review*, 20(1), 23–27.

- [3] Boulton, A. (2020). Data-driven learning for younger learners: Obstacles and optimism. In P. Crosthwaite (Ed.), *Data-driven learning for the next generation: Corpora and DDL for pre-tertiary learners* (pp. xiv–xx). Routledge.
- [4] Braun, S. (2007). Integrating corpus work into secondary education: From data-driven learning to needs-driven corpora. *ReCALL*, 19(3), 307–328.
<https://doi.org/10.1017/S0958344007000535>
- [5] Crosthwaite, P. (Ed.). (2020). *Data-driven learning for the next generation: Corpora and DDL for pre-tertiary learners*. Routledge.
<https://doi.org/10.4324/9780429425899>
- [6] Crosthwaite, P., & Baisa, V. (2023). Generative AI and the end of corpus-assisted data-driven learning? Not so fast! *Applied Corpus Linguistics*, 3,
<https://doi.org/10.1016/j.acorp.2023.100066>
- [7] Flowerdew, J. (2024). Data-driven learning: From Collins Cobuild Dictionary to ChatGPT. *Language Teaching*, 1–18.
- [8] Flowerdew, L. (2025). Constructivist Approaches to Data-Driven Learning. In: McCallum, L., Tafazoli, D. (eds) *The Palgrave Encyclopedia of Computer-Assisted Language Learning*. Palgrave Macmillan, Cham. https://doi.org/10.1007/978-3-031-51447-0_53-1
- [9] Gatto, M. (2014). *The Web as Corpus: Theory and practice*. Bloomsbury Academic.
- [10] Harel, I., & Papert, S. (1990). Software design as a learning environment. *Interactive Learning Environments*, 1(1), 1–32.
<https://doi.org/10.1080/1049482900010102>
- [11] Hirata, E. (2025). Young learners and data-driven learning. In L. McCallum & D. Tafazoli (Eds.), *The Palgrave encyclopedia of computer-assisted language learning*. Palgrave Macmillan.
https://doi.org/10.1007/978-3-031-51447-0_49-1
- [12] Johns, T. (1990). From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning. *CALL Austria*, 10, 14–24.
- [13] Kilgariff, A. (2007). Googleology is bad science. *Computational Linguistics*, 33, 147–151.
- [14] Kilgariff, A., Reddy, S., Pomikálek, J., & Avinesh, P. V. S. (2010a). A corpus factory for many languages. In *LREC workshop on Web Services and Processing Pipelines*, Malta.

- [15] Kilgarriff, A., Kovář, V., Krek, S., Srdanovic, I., & Tiberius, C. (2010b). A quantitative evaluation of word sketches. In *Proceedings of the 14th EURALEX International Congress* (pp. 372–379). The Netherlands.
- [16] Lin, P. (2023). ChatGPT: Friend or foe (to corpus linguists)? *Applied Corpus Linguistics*, 3(3), Article 100065. <https://doi.org/10.1016/j.acorp.2023.100065>
- [17] Ljubešić, N. & Klubička, F. (2014). {bs, hr, sr} WaC – Web corpora of Bosnian, Croatian
- [18] and Serbian. In F. Bildhauer & R. Schäfer (Eds.), *Proceedings of the 9th Web as Corpus Workshop (WaC-9) @ EACL 2014* (pp. 29-35). Gothenburg, Sweden: Association for Computational Linguistics.
- [19] Mikelić Preradović, N., Posavec, K., & Unić, D. (2019). Corpus-supported foreign language teaching of less commonly taught languages. *International Journal of Instruction*, 12(4), 335–352.
- [20] Miličević, M., & Samardžić, T. (2010). Korpusi kao sredstvo za postizanje autonomije u učenju stranog jezika. In J. Vučo & B. Milatović (Eds.), *Autonomija učenika i nastavnika u nastavi jezika i književnosti* (pp. 387–399). Filozofski fakultet.
- [21] Moderc, S., Stanković, R., Tomašević, A., & Škorić, M. (2023). An Italian-Serbian sentence aligned parallel literary corpus. *Review of the NCD*, 43, 78–91. <https://www.ncd.matf.bg.ac.rs/issues/43/09.pdf>
- [22] Perišić, O. (2025). Corpora vs. Web nell'apprendimento del serbo come lingua straniera. In B. Avšar, F. Brachini, V. Cavalloro et al. (Eds.), *Innovamenti: Spazi e percorsi per una ricerca multidisciplinare* (pp. 341–350). Edizioni Università per Stranieri di Siena.
- [23] Perišić, O. (2023). *Il corpus per imparare il serbo. Il futuro dell'apprendimento linguistico*. Edizioni dell'Orso.
- [24] Perišić, O. (2020). Collocazioni nella didattica del serbo come lingua straniera. *Filolog*, 11(22), 88–115.
- [25] Perišić Arsić, O. (2018). Mogućnost primene korpusa u nastavi srpskog jezika kao stranog. In M. Kovačević (Ed.), *Savremena proučavanja jezika i književnosti* (pp. 187–199). Filološko-umetnički fakultet u Kragujevcu.
- [26] Perišić Arsić, O. (2020). Translating lexical gaps: A contrastive corpus-based analysis. In M. Matešić & A. Memišević (Eds.), *Language and mind: Proceedings from the 32nd International Conference of the Croatian Applied Linguistics Society* (pp. 93–108). Peter Lang.
- [27] Perišić, O., Stanković, R., Ikonić Nešić, M., & Škorić, M. (2023). It-Sr-NER: CLARIN compatible NER and geoparsing web services for Italian and Serbian parallel text. In T. Erjavec & M. Eskevich (Eds.), *Selected papers from the*

- CLARIN Annual Conference 2022 (pp. 99–110). Linköping University Electronic Press. <https://doi.org/10.3384/ecp198>
- [28] Posavec, K. (2018). Uporaba korpusa u poučavanju hrvatskoga kao drugoga i inoga jezika. *Studia Lexicographica*, 12(22), 63–84.
- [29] Sinclair, J. McH. (2000). Lexical grammar. *Naujoji Metodologija*, 24, 191–203.
- [30] Stewart, D. (2025). Lost in the forest: Flashpoints for language learners in Sketch Engine. *Iperstoria*, 25, 537–554. <https://doi.org/10.13136/2281-4582/2025.i25.1585>
- [31] Tekwa, K. (2024). Artificial Intelligence, Corpora, and Translation Studies. In D. Li & J. Corbett (Eds.), *The Routledge Handbook of Corpus Translation Studies* (pp. 103–118).
- [32] Vitaz, M. & Poletanović, M. (2019). Data-Driven Learning. The Serbian case. *EL.LE*, 8(2), 409–422.
- [33] Zhang, M. & Crosthwaite, P. (2025). More human than human? Differences in lexis and collocation within academic essays produced by ChatGPT-3.5 and human L2 writers. *International Review of Applied Linguistics in Language Teaching*. <https://doi.org/10.1515/iral-2024-0196>
- [34] Zhang, R., & Zou, D. (2024). Self-regulated second language learning: A review of types and benefits of strategies, modes of teacher support, and pedagogical implications. *Computer Assisted Language Learning*, 37(4), 720–765. <https://doi.org/10.1080/09588221.2022.2055081>

Corpora for Learning Serbian as a Foreign Language in the AI Era

Olja Perišić

Summary

This study explores the challenges of corpus-based teaching of Serbian as a Foreign Language in the context of artificial intelligence technologies. While there are ongoing global discussions about the future of DDL in the AI era, the Serbian context demonstrates a disconnect between technological advancements, particularly those developed by IT experts from JeRTeh, and their application in foreign language teaching.

The research focuses on specific linguistic areas that have been examined in previous corpus studies: polysemy (*kuća–dom*), lexical gaps (*babine*) and positive or negative word connotations (*žena*). The findings show that, in tasks such as distinguishing between closely related

synonyms, corpus searches offer more precise and reliable results than AI, due to their ability to provide frequency-based data. Generative models are useful for lexical gaps, particularly for beginner learners, though corpus-based methods yield strong results when sufficient technical knowledge is available. However, when analysing word connotations, generative models demonstrate significant limitations, frequently avoiding negative content due to embedded safety filters, rendering them unsuitable for discourse analysis.

The main issue is that corpora are not sufficiently integrated into foreign language teaching in Serbia, which limits the ability to make comprehensive comparisons with AI tools. The study recommends training more teachers in corpus-based methods and promoting integrative approaches that combine corpora and generative models. It also advocates cultivating critical thinking, which is central to DDL and involves learner engagement with search engines and AI tools, as well as corpora.

Ultimately, this research offers a model for the thoughtful incorporation of AI into language learning, particularly for less commonly taught languages such as Serbian, while preserving the strengths of corpus-based methodology.

Keywords: corpora, Data Driven Learning, Serbian as a foreign language, JeRTeh, artificial intelligence, generative models.

Named Entity Recognition for Pre-modern Serbian: A Preliminary Study

Scientific paper

DOI: 10.18485/judig.2025.1.ch17

*Marija Đokić Petrović*¹,  0000-0002-0568-5219

*Mihailo St. Popović*²,  0000-0002-3128-2210

*Vladimir Polomac*³,  0000-0001-5140-9330

Abstract

Digital Humanities represent a significant advance in the research and preservation of cultural heritage, enabling the application of computational methods to the analysis of historical documents. In this context, Named Entity Recognition (NER) stands out as an important technique that allows scholars to automatically extract, classify, and organize essential information from historical sources. In this study, we focus on developing an NER model tailored to documents written in pre-modern Serbian. A NER model was trained on the historical corpus that included the Banjska Chrysobull, the third example of Dečani Chrysobull, and the collection of 13th-century charters and letters from the Dubrovnik archive. The model was evaluated on two representative documents: the Will of Miloš Belmužević from 1500 and unpublished manuscript from 1778 preserved in the archive of the Greek Orthodox Church of St. George in Vienna. The developed NER model is based on a Convolutional Neural Network architecture and trained to recognize three distinct entity types—personal names, toponyms, and demographic references. Although the model achieved an F1 score of ≈ 0.92 on the development set, its performance declined on the evaluation set, i.e., previously unseen

¹ School of Computing, Union University, Kneza Mihaila 6/VI, 11000 Belgrade, Serbia, mdjokicpetrovic@raf.rs

² Austrian Academy of Sciences, Institute for Medieval Research, Georg-Coch-Platz 2, 1010, Vienna, Austria, mihailo.popovic@oeaw.ac.at

³ University of Kragujevac, Faculty of Philology and Arts, Liceja Kneževine Srbije 1A, 34000, Kragujevac, Serbia, v.polomac@filum.kg.ac.rs

documents. Generally, this study presents the first application of NER to medieval Serbian documents written in pre-modern Serbian, thereby contributing to historical research and the preservation of cultural heritage. The proposed approach establishes a good foundation for further refinement of the developed model.

Keywords: Named Entity Recognition; Pre-modern Serbian; Digital Humanities; Cultural Heritage

1. Introduction

In recent decades, the field of Digital Humanities (DH) has emerged as a transformative approach to the study, interpretation, and preservation of cultural heritage (Kagaba Amina, 2024). This interdisciplinary field has proven particularly valuable in the context of historical documents (Milligan, 2022). Specifically, documents written in archaic scripts benefit greatly from digital techniques that enhance accessibility and deepen scholarly understanding (Alstola & Svard, 2024).

Among the various techniques employed in DH, Named Entity Recognition (NER) plays a particularly important role. It is designed to identify and classify entities in unstructured text into predefined categories such as persons, locations, organizations, demographics, events, and others (Jehangir, Radhakrishnan, & Agarwal, 2023). In the field of history, NER serves as a powerful tool for extracting knowledge from historical documents, enabling deeper semantic analysis and supporting more comprehensive historical interpretation (Ehrmann, Hamdi, Pontes, Romanello, & Doucet, 2023).

In the context of the Serbian language, NER has been increasingly applied to modern and printed texts with the support of machine learning and transformer-based models (Marovac, Avdić, & Milošević, 2023), thus enabling more advanced information extraction from contemporary historical sources. However, considerably less attention has been directed toward historical varieties of the Serbian language. In that sense, the pre-modern Serbian, which was used in medieval administrative, legal, and ecclesiastical documents, has remained largely unexplored in computational NER research.

To address this gap, this study focuses on developing an NER model tailored to documents written in pre-modern Serbian. The model was trained on a historical corpus that includes the Charter of King Stefan Uroš II Milutin to the Monastery of St. Stefan in Banjska (Banjska Chrysobull) (Trifunović, 2011), the third example of the Charter of King Stefan Uroš III Nemanjić to the Monastery of Visoki Dečani (Dečani Chrysobull) (Ivić

& Grković, 1976), and the collection of charters and letters from the 13th-century preserved in the Dubrovnik archive (Stojanović, 1929). To evaluate the model's performance, we selected two representative documents: the Will of Miloš Belmužević (Polomac, 2025; Obradović, 2024) from 1500 and unpublished manuscript from 1778, housed in the archive of the Greek Orthodox Church of St. George in Vienna (Plöchl, 1983; Ransmayr, 2018; Turczynski, 1959; Seirinidou, 2011).

The developed NER model, employing a Convolutional Neural Network (CNN) architecture and trained to recognize three distinct entity types—personal names, toponyms, and demographic references—achieved an F1 score of approximately 0.92 on the development set. Its performance, however, declined when applied to previously unseen documents, reflecting the linguistic and orthographic variability inherent in historical sources. Despite these challenges, this study represents the first application of NER for pre-modern Serbian, thereby contributing to both historical scholarship and the preservation of cultural heritage. The presented approach provides a robust foundation for further methodological improvements of the developed NER model.

The remainder of this paper is structured as follows. Section 2 provides an overview of related work. Section 3 presents the development of the NER model and the evaluation strategy. Section 4 discusses the findings. Finally, Section 5 offers concluding remarks and outlines directions for future research.

2. Related Work

NER is a fundamental task in Natural Language Processing (NLP) that involves identifying and categorizing entities in unstructured text into predefined classes such as personal names, organizations, locations, time expressions, quantities, and other domain-specific types (Jehangir, Radhakrishnan, & Agarwal, 2023). Although NER has been an established NLP task since the 1990s, its application to the Serbian language remained limited until the early 2010s, when more systematic and domain-specific research efforts began to emerge (Marovac, Avdić, & Milošević, 2023).

Early efforts in Serbian-language NER include the system developed by (Vitas & Pavlović-Lažetić, 2008), which combined morphological and lexical analysis with dictionary-based resources to recognize personal names and geographic entities. Authors (Ljubešić, Stupar, Jurić, & Agić) developed a machine learning-based NER system for Croatian and Slovene languages, closely related to Serbian, using conditional random fields. The model was trained on annotated web and

news corpora (e.g., SETimes, Vjesnik) and incorporated linguistic and distributional similarity features derived from large unannotated monolingual corpora. Krstev, Obradović, Utvić, & Vitas (2014) enhanced an earlier rule-based approach (Krstev, 1997), incorporating transducers and thesauri to identify personal and geopolitical entities. This approach was later utilized to construct a gold-standard dataset of news articles annotated with personal names, which served as training data for machine learning models such as Stanford NER and spaCy (Šandrih, Krstev, & Stanković, 2019). In the context of cultural heritage, (Tanasijević, 2019) developed a system for annotating cultural heritage documents with metadata by recognizing entities like years, person names, and document topics. Authors in (Todorović Šandrih, Krstev, Stanković, & Nešić Ikonić, 2021), developed a NER model (SrpCANNER) based on a CNN architecture, trained on a corpus of Serbian novels written between 1840 and 1920. The model was designed to recognize seven named entity categories—persons, locations, organizations, roles, events, demos, and art works—and achieved an F1 score of approximately 91% on the test dataset. Authors in (Nešić Ikonić, Petalinkar, Stanković, & Utvić, 2024) introduced the SrpCNeL model, which builds upon the NER backbone SrpCANNER2 by incorporating an entity-linking layer designed to align recognized entities with entries in the Wikidata knowledge base. The SrpCANNER2 model is trained using the spaCy Python module, employing the same model architecture as SrpCANNER (Todorović Šandrih, Krstev, Stanković, & Nešić Ikonić, 2021) across a diverse dataset containing sentences from Serbian novels (1840-1920), legal documents, as well as sentences generated from the Wikidata knowledge base and the Leximirka lexical database.

More recently, transformer-based models have brought significant advancements to NER in South Slavic languages. For example, (Ljubešić & Lauc, 2021) introduced a multilingual transformer model BERTiĆ, pre-trained on 8 billion tokens from web-crawled corpora in Bosnian, Croatian, Montenegrin, and Serbian. This model was fine-tuned on several domain-specific datasets, including SETimes.SR (Batanovic, Ljubešić, & Samaradžić, 2018), ReLDI-sr (Ljubešić, Erjavec, Batanović, Miličević, & Samardžić, 2025), and news article corpora, demonstrating strong performance in modern NER tasks. Authors in (Ikonić Nešić, Petalinkar, Škorić, & Stanković, 2024) conducted a comparison of different architectures and techniques for preparing NER models via integrating BERT with spaCy. They trained models to recognize seven entity types—persons, locations, organizations, professions, events, demonyms, and artworks—using a set of Serbian novels (1840-1920), modern newspaper

articles, and knowledge-base-derived sentences from the Wikidata knowledge base and Leximirka lexical database.

Despite the progress of NER for the contemporary Serbian language, the application of such methods to its historical varieties remains limited. This study presents the first attempt to develop and apply an NER model tailored to such script. By doing so, this paper addresses a critical gap in the digital processing of Serbian medieval textual heritage and contributes to the broader objective of enhancing access to historical content for both scholars and the wider public.

3. NER Model for Pre-modern Serbian

In this section, we present the training process of the NER model⁴ for pre-modern Serbian, followed by a detailed evaluation of its performance.

3.1. Training

We trained our model on the training corpus, which includes Banjska Chrysobull, Dečani Chrysobull and the collection of 13th-century charters and letters preserved in the Dubrovnik archive. We should mention that the Banjska Chrysobull, along with the Dečani Chrysobull, are the richest medieval sources of personal and geographical names (Grković, 1983; Loma, 2013).

The training corpus was first manually annotated, including entities like personal names (PERS), locations (LOC), and demographic references (DEMO).

To prepare the corpus for training, we first segmented the texts into sentences, resulting in a total of 6715 sentences (38900 words), of which 4112 contain named entities. Table 1 shows the distribution of entities in the corpus.

Table 1. Distribution of Entities in the Training Corpus

Entity	Total	Unique
PERS	7562	2490
LOC	1525	1275
DEMO	85	54

⁴ https://github.com/marijadjokic/ner_pre_modern_serbian

We then created annotations list as tuples sentence-list of entities. An example of such tuple is: „се҃моу оубо сѣѣ бѣи прѣстоѣ в десноу ю҃ѣ ꙗ҃вии се не ꙗ҃ко работѣнь нь ꙗ҃ко вѣка и гѣ всѣхъ долоу вѣкоушоу ю҃го плѣть бестрѣ҃ноу творе и приводе кѣ вѣоу и сего паче моисеа прославляе не бо ꙗ҃коже моиси задѣнаа бѣиа видѣ вь камени“, „entities“: [[171, 177, „PERSON“], [201, 206, „PERSON“]].

Afterwards, spaCy v3.7.5 allows us to specify a custom CNN architecture within a simple text file. Using the quick-start widget⁵, we set up the default settings. For our model, the language was Serbian, containing only the *ner* component. The model was trained on a CPU. We made specific adjustments to the default configuration:

- [components.tok2vec.model.encode]: The token-to-vector layer size was increased from 96 to 300 (the maximum recommended value).
- [components.tok2vec.model.encode]: The architecture used is HashEmbedCNN, with input and output width set to 300, 4 convolutional layers, 2,000 rows in hash embedding tables, a window size of 1 token for concatenation during convolutions, and no pretrained static vectors.

We then employed a Python library to randomly partition the loaded annotated data into two separate sets: a training set and a testing set, where 20% of the data is allocated to the test set, while the remaining 80% is used for training. This stratification ensures that the model can be trained on the majority of the data and subsequently evaluated on a held-out subset, providing an unbiased estimate of its performance.

Model training ended up after 15 epochs having 92.08%, 93.11% and 91.08% F1 score, precision (P), and recall (R), on the development set, respectively. The number of epochs is automatically generated.

3.2. Evaluation

Our evaluation dataset comprises two historical documents previously mentioned: the Will of Miloš Belmužević from the 1500s (Figure 1) and the document from the Greek Orthodox Church of St. George in Vienna from 1778 (Figure 2).

⁵ Quick-start spaCy3 widget, <https://spacy.io/usage/training#quickstart>

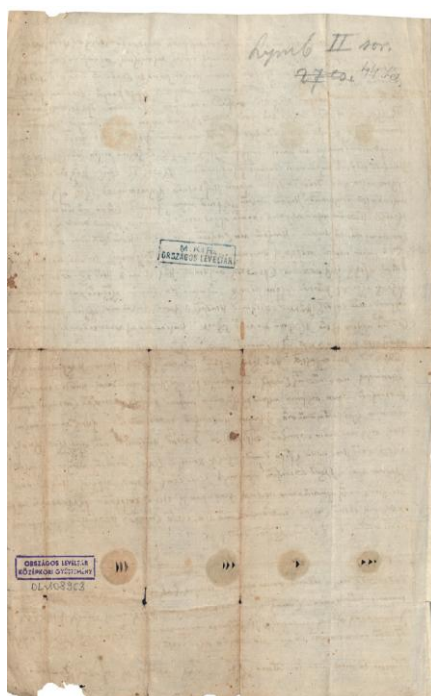
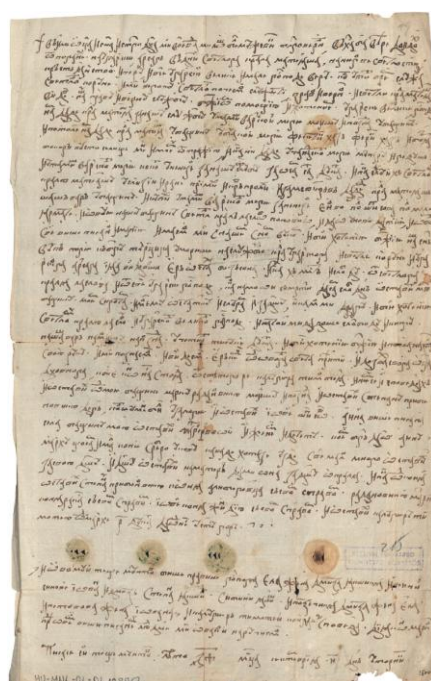


Figure 1. The Original Scanned Will of Miloš Belmužević, dated September 8, 1500. Source (Obradović, 2024)

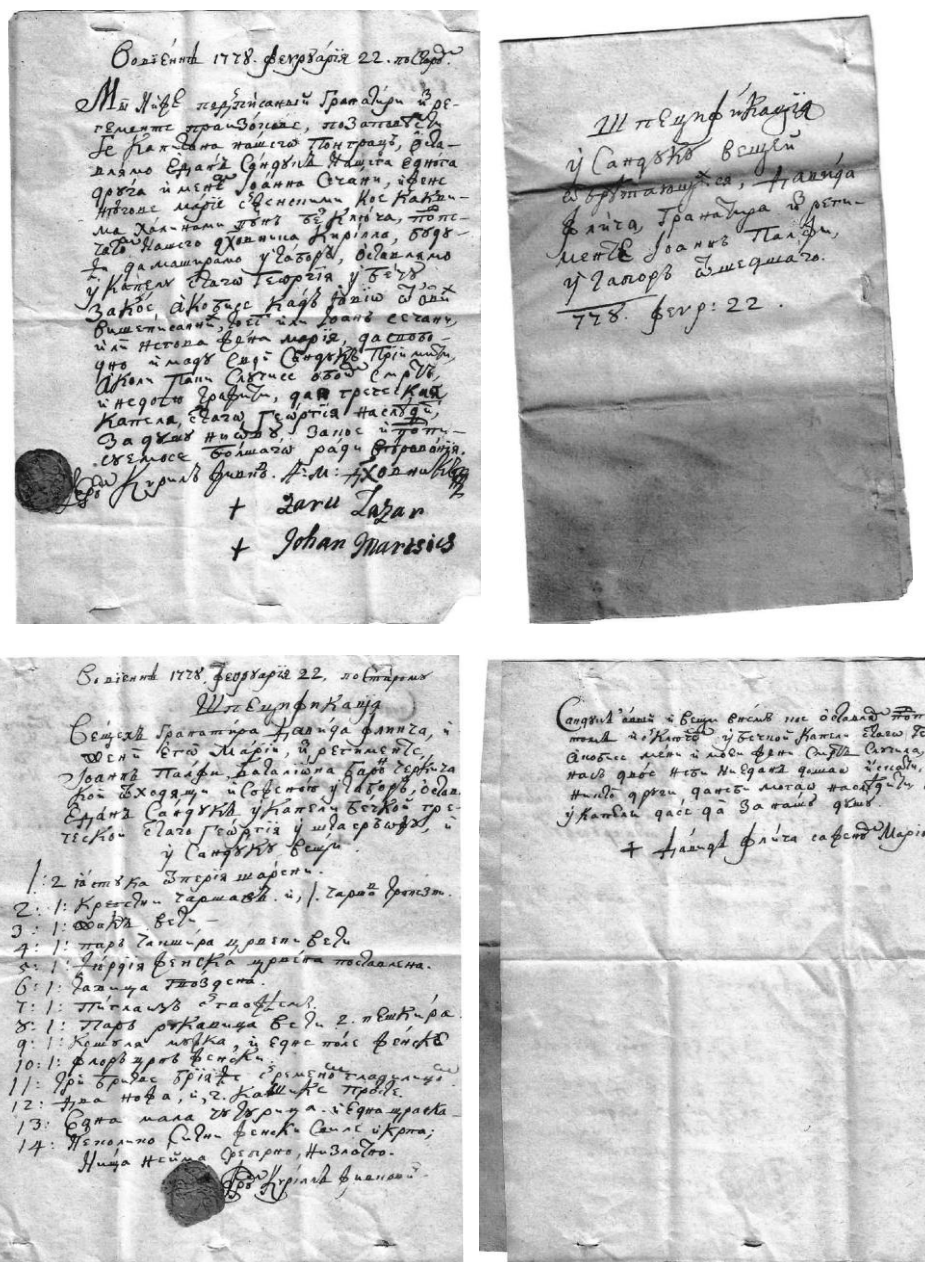


Figure 2. The Original Scanned Document from the Greek Orthodox Church of St. George in Vienna, dated February 22, 1778.

Table 2 provides the results, i.e. named entites of manual annotation of these documents.

Table 2. Distribution of Named Entities in the Evaluation Dataset

Entity	Will of Miloš Belmužević	Church Document
PERS	40	33
LOC	20	9
DEMO	2	0

We have run the previously trained model on these documents separately, and obtained the precision (P), recall (R) and F1 scores displayed in Table 3.

Table 3. Classification Report of NER Model on the Evaluation Dataset

Entity	Will of Miloš Belmužević			Church document		
	P	R	F1	P	R	F1
PERS	0.26	0.33	0.29	0.53	0.24	0.33
LOC	0.17	0.25	0.20	0.22	0.22	0.22
DEMO	0.0	0.0	0.0	0.0	0.0	0.0

Table 4 represents normalized confusion matrix, where the columns P_* (P_P , P_L , P_D) represent the predicted labels, while the rows T_* (T_P , T_L , T_D) represent the true (gold standard) labels (P refers to PERS, L refers to LOC, and D refers to DEMO). T_O refers to all tokens that do not belong to any of the annotated entity types, while P_O refers to all tokens that the model predicted as non-entity.

Table 4. Confusion Matrix on the Evaluation Dataset

Entity	Will of Miloš Belmužević				Church Document			
	P_P	P_L	P_D	P_O	P_P	P_L	P_D	P_O
T_P	13	1	0	26	8	0	0	25
T_L	3	5	0	12	0	2	0	7
T_D	0	1	0	1	0	0	0	0
T_O	34	0	0	2	7	7	0	0

4. Discussion

During the model development, we encountered 68 errors, primarily resulting from orthographic variation, tokenization mismatch, and annotation inconsistency. A major linguistic challenge stems from pre-modern orthography, which exhibits unstable spelling conventions and variable use of diacritical marks (e.g., ^а, ^ѣ, ^ѣ, ^ѣ, ...). These forms disrupt morphological segmentation, as the same morpheme may appear in several orthographic variants that do not align with modern normalization rules. Such discrepancies hinder both tokenization and the recognition of entity boundaries because standard NLP pipelines are trained on normalized modern orthography and lack mechanisms to account for the historical graphemic inventory of early Cyrillic scripts. Future improvements should thus include orthography-aware tokenization and morphological normalization layers capable of mapping graphemic variants to canonical lemma forms before NER training. For example, the errors can be observed in the following sentences:

[28, 31] по милости бжїеи ѣ стефанъ влѣславъ пишѣ те кнеже дѣбровъчський всѣи вѣпкинѣ (wrongly annotated entity влѣславъ);

[7, 16] мил'ко богдановиѣ а брат муѣ годю и драгославъ а снѣ моу прибиѣ а дѣд моу братохна (wrongly annotated entity богдановиѣ);

[17, 19] ѣ гѣ бѣ наѣго иѣ хѣ (wrongly annotated entity иѣ хѣ)

Following manual correction of the misaligned entity spans, the model achieved robust results on the development set. However, the model's performance on previously unseen documents remained limited, as indicated by the metrics in Table 3 and Table 4. A detailed study was therefore conducted. It is important to note that both the training and validation corpora were fully lower-cased.

In the case of the document from the Greek Orthodox Church, the model is significantly more effective at identifying PERS than LOC entities. It correctly detected given names such as марїе, марїя, zaru, johan, martsis, марїи, реуриѣ, and реуриѣ. Seven false positives were recorded, with the most notable being the word капитана ("captain") and нѣгове ("his"), which the model misclassified as personal names. For LOC entities, two true positives were retrieved (вїеннѣ and капели бечкой), whereas seven false positives appear - including печатѣ, сандѣкъ and штаерѣѣѣ. The last token was rejected as a true positive solely because of the trailing comma. No DEMO-type entities occurred in this document.

In the case of the Will of Miloš Belmužević, the model exhibited a similar pattern: higher recall for personal names than for toponyms. The model

correctly recognised forms such as милѡ^щ, вѹко^м, матеѡшь, тимотеи, маркѹ, прибенѡвику, дмитра, ѡкишкѡ, дмитрь, степа^н, ела, ѡвана, and вѹкѡ. Nevertheless, thirteen false positives were annotated. The most illustrative error was the token бѣмѹжеви^к, which was mis-labelled because its non-standard characters prevented normalisation; the same mechanism explains the false assignments for маркѹ and матеѡ. Tokens such as ма^трѣ (“mother”) and неговѹ (“his”) were likewise mis-tagged as names. For LOC entities, the model proposed моишѹ, мѹнарѹ, ѣнѡ, ша^варь, and ко^тко, but confusion between PERS and LOC entities was common. In addition, two DEMO-type entities appear as false negatives.

In general, the model performed well on the development set but still struggled with previously unseen material. Across the evaluation dataset, the model identified personal names more reliably than locations, which is expected, given the large number of PERS entities in the training data. Findings underscore the need for domain-specific preprocessing, such as Unicode normalization and character-trained tokenization, as well as stringent annotation guidelines, to improve generalization and robustness on previously unseen documents.

It should be noted that, although the current evaluation corpus is relatively small and therefore insufficient for broader statistical generalization, the results are nonetheless valuable as a proof of concept demonstrating the feasibility of automatic entity recognition in historical Serbian texts. This initial success establishes a solid foundation for subsequent linguistic and computational advancements.

5. Conclusion and Future Work

This paper represents the first end-to-end NER pipeline designed for pre-modern Serbian. By assembling and annotating a historical corpus comprising of the Banjska Chrysobull, Dečani Chrysobull and the 13th-century charters and letters from the Dubrovnik archive, we trained a SpaCy-based model that achieves F1 $\approx$ 92%. Evaluation of two previously unseen documents—the document from the Greek Orthodox Church of St. George in Vienna (1778) and the Will of Miloš Belmužević (1500)—confirmed that the system substantially outperforms manual lookup but still struggles when confronted with non-standard orthography.

Despite these challenges, we could conclude that the proposed model enables faster analysis of Serbian medieval documents, thereby advancing historical scholarship while simultaneously enriching digital humanities methodologies.

Future work will focus on expanding the training corpus, particularly to increase the number of entities related to location and demographic data to improve model performance. Additionally, we will stringent annotation guidelines and explore how transformer-based models can be utilized for the NER task in the same domain.

References

- [1] Alstola, T., & Svard, S. (2024). Digital Humanities Meets Ancient Languages. U M. Nissinen, & J. Jokiranta (Urednici), *Changes in Sacred Texts and Traditions: Methodological Encounters and Debates* (str. 193-233). Society of Biblical Literature. Preuzeto sa <https://cart.sbl-site.org/books/0603116P>
- [2] Batanovic, V., Ljubešić, N., & Samaradžić, T. (2018). SETimes. SR--a Reference Training Corpus of Serbian. *Proceedings of the Conference on Language Technologies Digital Humanities 2018 (JT-DH 2018)*, (str. 11-17).
- [3] Ehrmann, M., Hamdi, A., Pontes, E. L., Romanello, M., & Doucet, A. (2023). Named Entity Recognition and Classification on Historical Documents: A Survey. *ACM Computing Surveys*, 56(2), 1-47. doi:10.1145/3604931
- [4] Grković, M. (1983). *Imena u dečanskim hrisovuljama*. Novi Sad: Filozofski fakultet u Novom Sadu, Institut za južnoslovenske jezike.
- [5] Ikonić Nešić, M., Petalinkar, S., Škorić, M., & Stanković, R. (2024). BERT downstream task analysis: Named Entity Recognition in Serbian. *Conference on Information Technology and its Applications* (str. 333-347). Cham: Springer Nature Switzerland.
- [6] Ivić, P., & Grković, M. (1976). *Dečanske hrisovulje*. Novi Sad: Institut za lingvistiku.
- [7] Jehangir, B., Radhakrishnan, S., & Agarwal, R. (2023). A survey on named entity recognition—datasets, tools, and methodologies. *Natural Language Processing Journal*, 3, 100017. doi:10.1016/j.nlp.2023.100017
- [8] Kagaba Amina, G. (2024). Digital Humanities: Using Technology to Analyze Cultural Artifacts. *IDOSR Journal of Humanities and Social Sciences*, 9(3), 1-8. doi:10.59298/IDOSRJHSS/2024/93180000
- [9] Krstev, C. (1997). *Jedan prilaz informatičkom modeliranju teksta i algoritmi njegove transformacije*. Retrieved from <http://elibrary.matf.bg.ac.rs/handle/123456789/4134>
- [10] Krstev, C., Obradović, I., Utvić, M., & Vitas, D. (2014). A System for Named Entity Recognition Based on Local Grammar. *Journal of Logic and Computation*, 24(2), 473-489. doi:10.1093/logcom/exs079
- [11] Loma, A. (2013). *Toponomija Banjske hrisovulje*. Beograd: Srpska akademija nauke i umetnosti.

- [12] Ljubešić, N., & Lauc, D. (2021). Bertić—the Transformer Language Model for Bosnian, Croatian, Montenegrin and Serbian. *arXiv preprint arXiv:2104.09243*.
- [13] Ljubešić, N., Erjavec, T., Batanović, V., Miličević, M., & Samardžić, T. (2025). *ReLDI-NormTagNER-sr 2.1*. Preuzeto sa ReLDI CENTAR ZA JEZIČKE PODATKE: <https://github.com/reldi-data>
- [14] Ljubešić, N., Stupar, M., Jurić, T., & Agić, Ž. (2013). Combining Available Datasets for Building Named Entity Recognition Models of Croatian and Slovene. *Slovenščina 2.0: empirical, applied and interdisciplinary research*, 1(5), 35-57.
- [15] Marovac, U., Avdić, A., & Milošević, N. (2023). A survey of resources and methods for natural language processing of serbian language. *arXiv preprint arXiv:2304.05468*. doi:<https://arxiv.org/abs/2304.05468>
- [16] Milligan, I. (2022). *The transformation of historical research in the digital age*. Cambridge University Press. doi:10.1017/9781009026055
- [17] Nešić Ikonić, M., Petalinkar, S., Stanković, R., & Utvić, M. a. (2024). SrpCNNeL: Serbian Model for Named Entity Linking. *19th Conference on Computer Science and Intelligence Systems (FedCSIS)* (str. 465-473). IEEE.
- [18] Obradović, N. (2024). Testament Miloša Belmuževića. *Mešovita grada*, 45, 35-56. doi:10.34298/IC2473035O
- [19] Plöchl, W. M. (1983). Die Wiener orthodoxen Griechen. Eine Studie zur Rechts- und Kulturgeschichte der Kirchengemeinden zum Hl. Georg und zur Hl. Dreifaltigkeit und zur Errichtung der Metropolis von Austria. *Kirche und Recht*. Preuzeto sa <https://ixtheo.de/Record/013109626>
- [20] Polomac, V. (2025). Srpski jezik u Ugraskoj krajem XV veka (na primeru testamenta vojvode Miloša Belmuževića). (M. Kovačević, Ur.) *Aktuelna pitanja istorije srpskog jezika*. Preuzeto sa u štampi
- [21] Ransmayr, A. (2018). *Untertanen des Sultans oder des Kaisers: Struktur und Organisationsformen der beiden Wiener griechischen Gemeinden von den Anfängen im 18. Jahrhundert bis 1918*. Vienna : V&R unipress, Vienna University Press. Preuzeto sa <https://ucrisportal.univie.ac.at/de/publications/untertanen-des-sultans-oder-des-kaisers-struktur-und-organisation>
- [22] Seirinidou, V. (2011). *Έλληνες στη Βιέννη (18ος – μέσο 19ου αιώνα)*[=Greeks in Vienna, 18th - mid 19th C.]. Αθήνα: Ηρόδοτος.
- [23] Stojanović, L. (1929). *Stare srpske povelje i pisma* (T. 1). Beograd: Srpska kraljevska akademija.
- [24] Šandrih, B., Krstev, C., & Stanković, R. (2019). Development and Evaluation of Three Named Entity Recognition Systems for Serbian-the Case of Personal Names. U R. Mitkov, & G. Angelova (Ur.), *Proceedings of the International*

Conference on Recent Advances in Natural Language Processing (RANLP 2019) (str. 1060-1068). Varna, Bulgaria: INCOMA Ltd. doi:10.26615/978-954-452-056-4_122

- [25] Tanasijević, I. (2019). Toward Automatic Tagging of Cultural Heritage Documents. *IPSI Transactions on Advanced Research, TAR*, 15(1).
- [26] Todorović Šandrih, B., Krstev, C., Stanković, R., & Nešić Ikonić, M. (2021). Serbian NER&Beyond: The Archaic and the Modern Intertwined. U R. Mitkov, & G. Angelova (Ur.), *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)* (str. 1252-1260). Held Online: INCOMA Ltd. Preuzeto sa <https://aclanthology.org/2021.ranlp-1.141/>
- [27] Trifunović, Đ. (2011). Charter of King Milutin to the Banjska Monastery : The Chrysobull of St. Stephen. Službeni glasnik, Srbija. Retrieved from <https://books.google.at/books?id=ZjQeMwEACAAJ>
- [28] Turczynski, E. (1959). Die Deutsch-Griechischen Kulturbeziehungen Bis Zur Berufung KöNig Ottos. (F. Valjavec, Ur.) *Südosteuropäische Arbeiten*, 48.
- [29] Vitas, D., & Pavlović-Lažetić, G. (2008). Resources and Methods for Named Entity Recognition in Serbian. *INFOtheca - Journal of Informatics & Librarianship*, 9(1-2), 35a.

Prepoznavanje imenovanih entiteta za istorijske varijetete srpskog jezika: preliminarna studija

Marija Đokić Petrović,
Mihailo St. Popović,
Vladimir Polomac

U ovom radu predstavljen je prvi model za prepoznavanje imenovanih entiteta za istorijske varijetete srpskog jezika. Model je treniran na ručno anotiranom istorijskom korpusu koji obuhvata povelju Kralja Uroša II manastiru Banjska (Banjska hrisovulja), trećem primerku povelje Kralja Stefana Uroša III Nemanjića manastiru Visoki Dečani (Dečanska hrisovulja) i skupu povelja i pisama iz dubrovačkog arhiva iz XIII veka. Anotacije su obuhvatile tri kategorije entiteta – lična imena (PERS), toponime (LOC) i demografske reference (DEMO) – pri čemu su ukupno zabeležena 9172 entiteta u 6715 rečenica. Konkretno, 7562 entiteta tipa PERS, 1525 entiteta tipa LOC i 85 entiteta tipa DEMO. Broj anotiranih rečenica je iznosio 4112.

Za obuku modela je primenjena Convolutional Neural Network arhitektura u okviru spaCy-ja. Posle 15 epoha model je na razvojnom skupu postigao preciznost (P) 93,11%, odziv (R) 91,08 % i F1 92,08%.

Evaluacija je izvršena nad dva dokumenta: Testamentu Miloša Belmuževića iz 1500. godine i do sada neobjavljenom dokumentu koji se nalazi u arhivi grčke pravoslavne crkve Svetog Đorđa u Beču, a koji datira iz 1778. godine. Na ovom setu performanse modela su opale: F1 rezultati za PERS, LOC i DEMO entitete variraju od 0% do 33%, uz izraženo bolju tačnost za lična imena nego za toponime.

Detaljna analiza 68 uočenih grešaka prilikom procesa treniranja modela, pokazala je da su najčešći uzroci lošeg rada modela ortografska kompleksnost i retke dijakritike koje remete tokenizaciju, što dovodi do neusklađenih opsega anotacija. Rešavanjem problema pogrešno anotiranih entiteta značajno je poboljšana stabilnost treninga i metrika na razvojnom skupu, ali su problemi ostali prisutni na tekstovima koji nisu poznati modelu.

Ovaj rad doprinosi digitalnim humanističkim istraživanjima i razvoju alata za automatsku obradu srpskih srednjovekovnih izvora, postavljajući osnovu za buduće studije. Nalazi ukazuju na potrebu za proširenjem korpusa, Unicode normalizacijom, karakter-baziranom tokenizacijom, strožim smernicama za anotaciju i primenom transformer modela radi bolje generalizacije nepoznatih dokumenata.

Ključne reči: prepoznavanje imenovanih entiteta; istorijski varijeteti srpskog jezika; digitalne humanističke nauke; kulturno nasleđe.

Frekvencija značenja pridjeva u hrvatskome medicinskom nazivlju – korpusno utemeljena analiza kolokacija

Naučni rad

DOI: 10.18485/judig.2025.1.ch18

Martina Pavić¹,  0000-0001-6061-9495

Apstrakt

Pridjevi u medicinskome području imaju ključnu ulogu u klasifikaciji i specifikiranju bolesti, dijagnoza, simptoma, indikacija i drugih pojmovnih kategorija. Proučavanje frekvencije njihovih značenja otkriva kako medicinski jezik konceptualizira i kategorizira zdravstvene pojave, primjerice prema intenzitetu (*blag, težak*), lokaciji (*desni, lijevi*) ili uzrocima (*infektivni, genetski*).

U ovome radu istražena je zastupljenost značenja 50 pridjeva s više od 100 pojava u znanstvenome korpusu suvremenih medicinskih časopisa veličine 5.318,950 riječi, koji je sastavljen u alatu *Sketch Engine* (Pavić, 2022). U analizu su uključeni i pridjevi iz općega jezika (npr. *desni, dubok, središnji*) koji se pojavljuju u višerječnim medicinskim nazivima.

Značenjska analiza temelji se na modelu odozgo prema dolje (engl. *top-down approach*), počevši od vršnih kategorija naziva, kao što su bolest, simptom i postupak, koje su potom proširene u skladu s uvidima iz korpusa.

Pridjevi su klasificirani u tri skupine: 1) pridjevi koji mijenjaju kategoriju naziva i značenje, 2) pridjevi koji mijenjaju kategoriju naziva, ne i značenje te 3) pridjevi koji ne mijenjaju ni kategoriju naziva ni značenje.

Rezultati prikazuju najčešća značenja pridjeva u hrvatskome medicinskom nazivlju te broj značenja koje pojedini pridjev ostvaruje u tipičnim kolokacijama. Cilj je rada korpusnom metodom utvrditi dominantna značenja pridjeva i prikazati kako vršne kategorije naziva utječu na promjene značenja.

¹ Institut za hrvatski jezik, mpavic@ihjj.hr

Ključne riječi: pridjevi, značenjska analiza, hrvatsko medicinsko nazivlje, znanstveni korpus, kolokacije, vršne kategorije naziva

1. Uvod

Frekvencijska analiza značenja pridjeva u medicinskome jeziku otkriva obrasce konceptualizacije i kategorizacije ključnih pojmova te pruža uvid u načine na koje se opis stvarnosti oblikuje u stručnome medicinskom diskursu. U medicinskome nazivlju, u kojemu je terminološka preciznost presudna za učinkovitu komunikaciju, pridjevi imaju važnu klasifikacijsku i diferencirajuću ulogu značajno doprinoseći stvaranju značenjske specifičnosti i pojmovne organizacije u okviru vršnih kategorija naziva poput bolesti, simptoma i medicinskih postupaka.

Unatoč tomu frekvencija značenja pridjeva u hrvatskome medicinskom jeziku dosad nije sustavno istraživana, iako upravo poznavanje njihove upotrebe i značenjskoga raspona omogućuje dublji uvid u strukturu hrvatskoga medicinskog nazivlja te pridonosi terminološkoj standardizaciji. U ovome se radu stoga istražuju kontekstno uvjetovana značenja pridjeva u medicinskome korpusu, s posebnim naglaskom na njihovu učestalost, varijabilnost u kolokacijama i u odnosu na vršne kategorije naziva u okviru kojih se pojavljuju. Kako bi se dobio pouzdan uvid u obrasce značenja pridjeva u hrvatskome medicinskom nazivlju, rad se temelji na sljedećim istraživačkim pitanjima:

1. Koja su najfrekventnija značenja pridjeva u hrvatskome medicinskom nazivlju?
2. Mijenja li se značenje pridjeva u različitim kolokacijama i u okviru različitih vršnih kategorija naziva?
3. Koje su vršne kategorije naziva najčešće u hrvatskome medicinskom jeziku?

Na temelju ovih pitanja postavljene su tri hipoteze:

- H1: Pridjevi u hrvatskome medicinskom nazivlju najčešće izražavaju vrijeme/trajanje (npr. *akutni*, *kronični*), uzrok (npr. *hormonski*, *virusni*) i lokaciju (npr. *periferni*, *središnji*).
- H2: Značenje pridjeva u medicinskome nazivlju mijenja se ovisno o kolokacijskome okruženju.
- H3: Najfrekventnije vršne kategorije u okviru kojih se pojavljuju medicinski pridjevi jesu ⟨bolest⟩, ⟨terapija⟩ i ⟨simptom⟩.

Prva hipoteza proizlazi iz temeljnih semantičkih domena medicinskoga mišljenja: kada, zašto i gdje, dok je druga hipoteza

utemeljena na konceptu kontekstne specifikacije značenja u terminološkim sustavima, prema kojemu značenje pridjeva varira ovisno o kolokacijama. Treća hipoteza polazi od stajališta da upravo kategorije ⟨bolest⟩, ⟨terapija⟩ i ⟨simptom⟩ čine osnovnu strukturu medicinskoga nazivlja te da pridjevi u tim kontekstima imaju klasifikacijsku funkciju.

Osnovni je cilj ovoga rada prikazati najfrekventnija značenja pridjeva u hrvatskome medicinskom nazivlju kako bi se stekao uvid u to što frekvencija može reći o načinu na koji se opisuju bolesti, stanja, dijagnoze, terapije, postupci i sl. Dodatni su ciljevi:

- utvrditi broj značenja za svaki pojedini pridjev
- istražiti koji pridjevi mijenjaju značenje u kolokacijama
- identificirati najčešće vršne kategorije naziva.

Rad je strukturiran u nekoliko poglavlja. Drugo poglavlje donosi teorijski okvir o pridjevima u specijaliziranome jeziku. U trećemu poglavlju prikazan je korpusni pristup te metodologija analize značenja i kolokacija. Četvrto poglavlje iznosi rezultate analize, uključujući klasifikaciju značenja, varijacije u kolokacijama i distribuciju po vršnim kategorijama naziva. Peto poglavlje sažima glavne uvide i donosi prijedloge za daljnja istraživanja. Prilog I sadržava tablicu pridjeva prema vršnim kategorijama naziva, kolokacijama, primjerima rečenica iz korpusa i značenjima, dok se u Prilogu II navode značenja svih 50 pridjeva pronađenih u korpusu.

2. Teorijski okvir

Brojna su lingvistička istraživanja posvećena medicinskom nazivlju (Lynch, 1997; Bowker, Hawkins, 2006; Gjuran-Coha, 2011; Gjuran Coha, Bosnar-Valković, 2013, 2017; Taylor, 2017), no pridjevi u tim analizama rijetko zauzimaju središnje mjesto. Imenice su tradicionalno u fokusu terminoloških istraživanja jer se smatraju temeljnim nositeljima značenja i referentima stabilnih pojmovnih jedinica u stručnome diskursu. Nasuprot tomu pridjevima je pripisana sekundarna, uglavnom modifikatorska funkcija, što je rezultiralo njihovim marginaliziranjem u proučavanju jezika struke.

Takav pristup potvrđuju i podatci o sastavu stručnih rječnika: L'Homme (2003) navodi da imenice i imenske fraze čine između 84,25 % i 97,5 % svih natuknica u rječnicima različitih stručnih područja, dok je udio pridjeva znatno manji (npr. 11 % u medicinskom nazivlju), a glagoli se pojavljuju još rjeđe.

Tradicionalna usmjerenost terminoloških istraživanja na imenice kao temeljne nositelje značenja sve se više dovodi u pitanje u suvremenim teorijama koje prepoznaju dinamičnost diskursa i jezičnu varijabilnost u stvarnoj uporabi. Komunikacijska terminološka teorija (Cabré, 1999, 2003), sociokognitivna terminologija (Temmerman, 2000) i terminologija okvira (Faber, 2009) razvijaju modele u kojima se značenje ne promatra kao statička kategorija, nego kao dinamična funkcija konteksta u kojima sudjeluju dinamične strukture znanja koje obuhvaćaju procese i događaje. Prema tim pristupima terminološka vrijednost leksičkih jedinica nije unaprijed zadana, nego se aktivira u specijaliziranim diskursima te se u tom okviru tzv. predikativne jedinice – glagoli, pridjevi, prilozi, ali i neke imenice – sve češće promatraju kao jezični izrazi temeljnih pojmova u stručnim domenama (L'Homme, 1998; Pimentel, 2012; L'Homme & Robichaud, 2014).

U skladu s time, iako se rjeđe pojavljuju kao natuknice u stručnim rječnicima, pridjevi u specijaliziranim tekstovima imaju važnu ulogu u oblikovanju i preciziranju značenja. Konceptualna struktura stručnoga znanja temelji se na hijerarhijskim i paradigmatiskim odnosima među pojmovima, pri čemu semantički odnosi poput hiponimije i antonimije omogućuju razlikovanje i sistematizaciju značenja (Lyons, 1977; Cruse, 1986). U takvome okviru pridjevi preuzimaju klasifikacijsku funkciju i doprinose uspostavljanju pojmovne preciznosti i hijerarhijske organizacije stručnoga znanja. Primjeri poput *dobročudni tumor* i *zločudni tumor* jasno pokazuju da pridjev nosi terminološku vrijednost i omogućuje distinkciju unutar istoga semantičkog polja u medicinske nazivlje. Slični rezultati zabilježeni su i u drugim domenama, primjerice u botanici (Pitkänen-Heikkilä, 2015), turizmu (Durán-Muñoz, 2019) i znanosti o kršu (Grčić Simeunović, 2015, 2020, 2021), gdje se pridjevi istodobno rabe za taksonomsku preciznost i terminološku diferencijaciju.

Unatoč različitim teorijskim pristupima sve je izraženiji konsenzus o tome da se pridjevima pristupi kao leksičkim jedinicama s visokim semantičkim potencijalom u okviru višečlanih naziva. U takvim konstrukcijama pridjevi nadilaze modifikatorsku funkciju i postaju ključne sastavnice terminoloških jedinica koje pridonose standardizaciji stručnoga jezika. Istodobno, porast složenosti medicinskoga nazivlja dovodi do sve češće upotrebe višerječnih naziva koji mogu biti višeznačni ili nedovoljno jasni, što potencijalno otežava njihovo razumijevanje i primjenu u stručnome jeziku (Bowker & Hawkins, 2006). U takvim okolnostima pridjevi često preuzimaju ključnu ulogu u razgraničenju i preciziranju pojmova.

Teorijski pomaci popraćeni su i razvojem metodoloških pristupa, među kojima se posebno istaknula korpusna analiza jer omogućuje detaljan uvid u frekvencijsku pojavnost, kolokacijske obrasce i funkcionalnu ulogu terminoloških jedinica u stvarnome diskursu (Sager, 1990; Rafaelli, 2004). U okviru korpusne lingvistike sustavno se istražuje uporaba pridjeva u velikim jezičnim bazama, što omogućuje dublje razumijevanje njihove semantičke raznolikosti i doprinosa razvoju medicinskoga nazivlja. S razvojem korpusne lingvistike povećava se i preciznost u analizi stvarne uporabe terminoloških jedinica, uključujući semantičke i sintagmatske obrasce pridjeva. U tome kontekstu Bourigault & Slodzian (1999) uvode pojam tekstne terminologije (fr. *terminologie textuelle*) kako bi označili pristup koji se temelji na stvarnoj uporabi, a ne samo na definicijama i leksikografskim konvencijama.

Jedan od ključnih pojmova u okviru takvoga pristupa jest nazivnost (engl. *termhood*), koju Kageura i Umino (1996) definiraju kao „stupanj u kojemu je jezična jedinica povezana s pojmom specifičnim za područje” (usp. Warburton, 2021). Nazivnost, dakle, označava terminološku povezanost, a ne isključivo učestalost pojedine riječi u specijaliziranome jeziku. Važno je istaknuti da nazivnost nije binarna, nego kontinuirana kategorija – pojedine riječi mogu pokazivati više ili manje izraženu terminološku povezanost ovisno o kontekstu uporabe. Primjerice, pridjevi *benigni* i *maligni*² imaju visoku nazivnost u višerječnim nazivima *benigni tumor* i *maligni tumor* jer označavaju jasno definirane medicinske pojmove povezane s klasifikacijom tumora prema njihovim patološkim svojstvima. Međutim, u izrazima poput *benigne promjene* i *maligne promjene* nazivnost je niža jer se ne odnose na ustaljene medicinske nazive, nego na šire opisne izraze koji se mogu pojavljivati i izvan užega medicinskog konteksta (npr. u metaforičnoj uporabi u općemu jeziku). Iako pridjevi *benigni* i *maligni* i u tim primjerima zadržavaju određeni stupanj terminološke prepoznatljivosti, njihova se nazivnost smanjuje jer više ne označavaju specifičan pojam unutar jedne terminološke mreže, nego služe kao kontekstni deskriptori. Slično vrijedi i za pridjev *upalni*, koji u kolokacijama poput *upalna bolest* i *upalni proces* pokazuje visoku nazivnost s obzirom na to da upućuje na specifične patofiziološke procese i čini dio terminološkoga sustava medicinske struke. Nasuprot tomu sintagma *upalni znakovi* opisuje skup simptoma (npr. crvenilo, otekline, bol, toplina i gubitak funkcije), odnosno ne označuje strogo definiran medicinski pojam, nego opće obilježje i stoga pridjev ima nižu nazivnost.

² U korpusu su zabilježene i rjeđe inačice *benigan* i *maligan*, no u analizi se ne izdvajaju kao zasebne leme jer dijele značenjsku i terminološku vrijednost s inačicama *benigni* i *maligni*.

Visok stupanj nazivnosti pokazuju sinonimni oblici *inflamacijski* i *inflamatorni* u svezi *inflamacijski/inflamatorni proces*, koja je potvrđena u korpusu i označava ustaljenu inačicu naziva u stručnoj uporabi.

Takve varijacije u stupnju nazivnosti pridjeva odražavaju se i na njihov značenjski raspon, koji se u stvarnome diskursu očituje različitim kolokacijskim obrascima. Analizom tih kolokacija u korpusnome okruženju moguće je pratiti frekvenciju i distribuciju značenja pridjeva, čime se otkrivaju obrasci njihova terminološkoga ponašanja u medicinskoj domeni.

3. Korpusna metodologija

U ovom je radu primijenjen znanstveni korpus suvremenih medicinskih časopisa, izrađen i obrađen u alatu *Sketch Engine*³ (Kilgariff i dr., 2012). Korpus obuhvaća 5.318.395 riječi, a temelji se na referentnome korpusu *hrWaC 2.2*, označenom RFTTaggerom (Ljubešić i Erjavec 2011; Ljubešić i Klubička 2014), čime je osigurana visoka razina morfosintaktičke točnosti. Analiza je provedena poluautomatski uporabom opcija *Wordlist*, *Word Sketch* i *Concordance*, što je omogućilo analizu pridjeva u kontekstu najčešćih kolokacija i vršnih kategorija naziva. U korpus su uključeni tekstovi iz sedam relevantnih hrvatskih medicinskih časopisa: *Acta Medica Croatica*, *Liječnički vjesnik*, *Medicus*, *Infektološki glasnik*, *Cardiologia Croatica*, *Physioterapia Croatica* i *Paediatrica Croatica* (vidi Pavić, 2022: 61–62). Na Slici 1. prikazani su podatci o korpusu, među ostalim broj riječi, pojavnica i rečenica.

Preliminarnom analizom obuhvaćeno je 50 pridjeva s više od 100 pojavnica u korpusu, uključujući i pridjeve iz općega jezika koji se učestalo pojavljuju kao sastavnice višječlanih medicinskih naziva. Odabrani su oni pridjevi čije je značenje bilo prozirno i precizno određeno iz konteksta. Uporabom opcije *Wordlist* dobiven je popis tih pridjeva s brojem pojavnica većim od sto ($N > 100$), koji su prikazani u Tablici 1.

³ *Sketch Engine* omogućuje preciznu identifikaciju leksičkih obrazaca koje bi bilo teško ili nemoguće ručno analizirati u velikom korpusu. Digitalni alati poput *Sketch Enginea* znatno smanjuju vrijeme analize i omogućuju veću preciznost u proučavanju specijaliziranih pojmova.

ZNANSTVENI ČLANCI user/MARTINA1/znanstveni_clanci • created November 24, 2021 at 2:14:05 PM

MANAGE CORPUS MANAGE SUBCORPORA COMPARE CORPORA TEXT TYPE ANALYSIS

GENERAL INFO

Language: Croatian

CORPUS DESCRIPTION & BIBLIOGRAPHY ⓘ

TAGSET

WORD SKETCH GRAMMAR

TERM GRAMMAR

COUNTS ⓘ

Tokens	7,432,737
Words	5,318,395
Sentences	279,168
Documents	7

LEXICON SIZES ⓘ

word?	323,517
tag	777
lempos?	228,663
gender_lemma	214,112
lc ⓘ	275,154
lemma	206,507
g ⓘ	5
n ⓘ	4
c ⓘ	9
pos	13

Slika 1. Znanstveni korpus suvremenih medicinskih časopisa (Pavić, 2022)

Tablica 1. Popis 50 visokofrekventnih pridjeva iz znanstvenoga korpusa suvremenih medicinskih časopisa (Pavić, 2022) i broj njihovih pojavnica

veliki (N = 7363)	mali (N = 4406)	visoki (N = 4037)	kronični (N = 3826)	akutni (N = 3318)
pozitivni (N = 3169)	niski (N = 2474)	terapijski (N = 2454)	primarni (N = 2419)	negativni (N = 2080)
krvni (N = 2000)	upalni (N = 1925)	bubrežni (N = 1836)	kirurški (N = 1794)	humani (N = 1716)
srednji (N = 1674)	lokalni (N = 1596)	antimikrobni (N = 1472)	laboratorijski (N = 1375)	arterijski (N = 1335)
kardiovaskularni (N = 1251)	infektivni (N = 1168)	maligni (N = 1165)	koronarni (N = 1147)	limfni (N = 1126)
venski (N = 1107)	dijagnostički (N = 1102)	alergijski (N = 1036)	lijevi (N = 1034)	invazivni (N = 984)
virusni (N = 926)	rezistentni (N = 918)	blagi (N = 869)	periferni (N = 865)	desni (N = 840)
unutarnji (N = 753)	patogeni (N = 580)	probavni (N = 560)	masni (N = 537)	peroralni (N = 529)
poslijeoperacijski (N = 468)	intravenski (N = 400)	ishemijski (N = 381)	hormonski (N = 315)	difuzni (N = 311)
degenerativni (N = 264)	gnojni (N = 258)	perkutani (N = 197)	izvanbolnički (N = 185)	intraoperacijski (N = 154)

Cilj istraživanja bio je dobiti preliminarni uvid u kolokacijske obrasce, značenjska obilježja i distribuciju tih pridjeva u medicinskome diskursu, uz nastojanje da se postigne ravnoteža između širine uzorka i dubine analize.⁴

Polazište značenjske analize bila je tipologija na 14 značenja pridjeva u hrvatskome medicinskom nazivlju: funkcija, uzrok, vrijeme/trajanje, fizička značajka, sastav, stanje, svojstvo, lokacija, ishod, primjena, izvor, dob, intenzitet i relacija (Pavić, 2022). Navedena se klasifikacija temelji na tipologiji pridjeva koju je razvio Dixon (1982, 2004) te na strukturi terminološke baze znanja o znanosti okoliša EcoLexicon (2019). Budući da se značenje relacije može promatrati uopćeno, važno je istaknuti da je u ovome članku ta kategorija primijenjena za pridjeve koji označavaju konceptualnu ili funkcionalnu povezanost s imenicom. Primjerice, u svezi *arterijska hipertenzija* pridjev *arterijski* ne označuje lokaciju, nego tip hipertenzije povezane s arterijama (funkcionalno-konceptualna veza), dok u svezi *bubrežni transplantat* pridjev *bubrežni* označuje stvarno mjesto, pa ima značenje lokacije. Slično, *bubrežni bolesnik* označava osobu koja pati od bolesti bubrega, pri čemu *bubrežni* uspostavlja konceptualnu vezu s imenicom *bolesnik*, što upućuje na relacijsko značenje pridjeva.

Istraživanje je provedeno prema modelu odozgo prema dolje (engl. *top-down approach*), koji se temelji na hijerarhijskim odnosima unutar terminološkoga sustava. Model vršnih kategorija djelomično je preuzet od Taylora (2017) i uključuje vršne kategorije naziva ⟨bolest⟩, ⟨lijek⟩, ⟨simptom⟩, ⟨indikacija⟩, ⟨dijagnoza⟩ i ⟨postupak⟩, kojemu su dodane kategorije ⟨medicinsko osoblje⟩, ⟨bolesnik⟩, ⟨nuspojava⟩, ⟨dio tijela⟩, ⟨medicinska oprema⟩ (Pavić, 2022). U ovome radu Taylorova kategorija naziva ⟨lijek⟩ uključena je u širu kategoriju naziva ⟨terapija⟩ s obzirom na to da se terapija najčešće leksikalizira upravo imenicom *lijek* ili vrstama lijeka. Dodatno su uključene i vršne kategorije naziva ⟨ljudski organizam⟩ i ⟨mikroorganizam⟩.

Iako sama frekvencijska analiza značenja pridjeva nije zahtijevala daljnju hijerarhijsku raščlambu postojećih kategorija naziva (npr. podvrste bolesti, simptoma ili indikacija), pojavili su se određeni metodološki izazovi zbog višeznačnosti i kontekstne ovisnosti vršnih kategorija naziva. Stoga je za precizno određivanje vršne kategorije naziva bilo nužno navoditi rečenice iz korpusa u kojima se naziv pojavljuje. Takav pristup omogućuje transparentnost i dosljednost u kategorizaciji, ističući da se ista

⁴ Buduća istraživanja proširit će ovu analizu na širi skup pridjeva kako bi se potvrdile ili eventualno osporile uočene tendencije.

leksička jedinica može, ovisno o kontekstu, svrstati u različite vršne kategorije.

Analiza je utemeljena na utvrđivanju vršnih kategorija naziva jer one omogućuju kontekstualizaciju i klasifikaciju značenjskih funkcija pridjeva u specifičnim leksičkim vezama. Primjerice, u kategoriji naziva ⟨bolest⟩ pridjevi *akutni* i *kronični* rabe se za izražavanje vremena odnosno trajanja. Takva učestala upotreba pridjeva u određenim kategorijama naziva pridonosi prepoznavanju tipičnih značenjskih funkcija i pomaže u izgradnji terminoloških sveza koje osiguravaju dosljednost i jasnoću u jeziku struke. S druge strane značenjska klasifikacija pridjeva *primaran* kontekstno je uvjetovana: u svezi *primarna hipertenzija* pridjevom se definira vrstu bolesti pa ima značenje stanja, dok u svezi *primarna glavobolja* pridjev označuje početak odnosno nastanak bolesti pa ima vremensko značenje. Također, analiza frekvencije pridjeva u okviru kategorija naziva može poslužiti kao pokazatelj terminoloških promjena i leksičkih trendova u medicinske nazivlju. Primjerice, sve učestalija pojava pridjeva *eksperimentalni* u kategoriji naziva ⟨terapija⟩ može signalizirati nove pristupe i istraživačke pomake u medicinskoj praksi, što upućuje na dinamičnost samoga nazivlja.

4. Rezultati

Korpusnim istraživanjem utvrđene su tri skupine pridjeva prema tome u koliko se vršnih kategorija naziva pojavljuju te koliko značenja izriču:

1. pridjevi koji u kolokacijama mijenjaju vršnu kategoriju naziva i značenje
2. pridjevi koji mijenjaju vršnu kategoriju naziva, ne i značenje
3. pridjevi koji ne mijenjaju ni vršnu kategoriju naziva ni značenje.

Nisu potvrđena sva značenja od kojih se pošlo u frekvencijskoj analizi 50 pridjeva. Naime, u korpusu nisu pronađena značenja *izvor* i *dob*, no dodano je značenje *vrijednost* jer ga izražavaju pridjevi *negativni*, *niski*, *srednji*, *visoki*, koji tvore višerječne medicinske nazive. Time je u korpusu potvrđeno sljedećih 13 značenja pridjeva: *funkcija*, *uzrok*, *vrijeme/trajanje*, *fizička značajka*, *sastav*, *stanje*, *svojstvo*, *lokacija*, *ishod*, *primjena*, *intenzitet*, *relacija*, *vrijednost*.

Prema rezultatima korpusnoga istraživanja samo su tri pridjeva razvrstana u skupinu onih koji u kolokacijama ne mijenjaju ni vršnu kategoriju naziva ni značenje. To su *patogeni*, koji izražava značenje uzroka s imenicama *antitijela*, *bakterije*, *mikroorganizmi* u okviru vršne

kategorije naziva ⟨mikroorganizam⟩, *peroralni*, koji izriče značenje primjene s imenicama *glukokortikoid*, *kortikosteroid*, *pripravak* u okviru vršne kategorije naziva ⟨terapija⟩, te pridjev *masni* koji ima značenje sastava s imenicama *jetra*, *tkivo*, *stanica* u okviru vršne kategorije naziva ⟨dio tijela⟩. Primjeri su rečenica iz korpusa:

- (1) Većina ovakvih pneumonija nastaje aspiracijom **patogenih** mikroorganizama iz orofaringealnog sekreta.
- (2) Ako se bolesnikovo stanje poboljša, treba utvrditi najnižu dnevnu dozu **peroralnih** glukokortikoida uz koju će astma biti dobro kontrolirana.
- (3) Ono može utjecati na razvoj inzulinske neosjetljivosti, nastanak metaboličkog sindroma, a lučenjem adipokina, čimbenika stvorenih u **masnim** stanicama, doprinosi upalnim procesima, pa se i iz tog razloga pretilost smatra kroničnom upalnom bolešću.

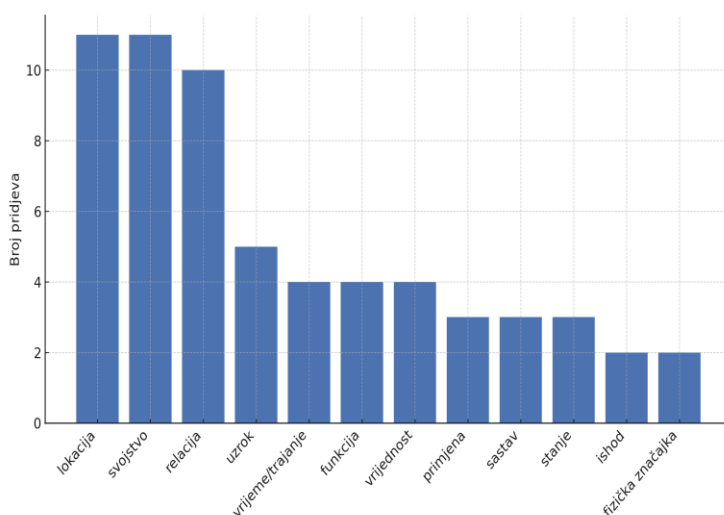
U skupinu pridjeva koji mijenjaju vršnu kategoriju naziva, ne i značenje ubrajaju se, među ostalim, pridjevi koji izražavaju vrijeme/trajanje (npr. *kronični* u kategorijama naziva ⟨bolest⟩, ⟨nuspojava⟩), lokaciju (npr. *periferni* u kategorijama naziva ⟨bolest⟩, ⟨ljudski organizam⟩), svojstvo (npr. *invazivni* u kategorijama naziva ⟨bolest⟩, ⟨postupak⟩) i uzrok (npr. *alergijski* u kategorijama naziva ⟨bolest⟩, ⟨terapija⟩).

- (3) Patohistološkim nalazom želučane sluznice utvrđen je umjereni **kronični** gastritis s blagom žarišnom atrofijom.
- (4) U razdoblju djetinjstva i adolescencije često se zbog **periferne** neuropatije javlja simptom akroparestezija koji se prezentira akutnom ili kroničnom boli u šakama i stopalima.
- (5) U većini opisanih slučajeva **invazivne** trihosporonoze (63%), u podlozi su bile maligne bolesti krvi kao glavni predisponirajući činitelj.
- (6) Kod jedne bolesnice zbog opsežne urtikarije i moguće **alergijske** reakcije na lijek odlučeno je primijeniti lijek u sljedećoj dozi uz individualnu procjenu doze kliničkog farmakologa.

U treću skupinu ulaze pridjevi koji mijenjaju vršnu kategoriju naziva i značenje. Primjerice, pridjev *infektivni* ostvaruje dva značenja: u kategoriji naziva ⟨bolest⟩ izražava uzrok (*infektivna mononukleoza*), dok u kategoriji naziva ⟨mikroorganizam⟩ ima značenje svojstva (*infektivni serovar*). Pridjev *srednji* označava lokaciju u vršnoj kategoriji naziva ⟨dio tijela⟩ (*srednje uho*), dok u kategoriji ⟨ljudski organizam⟩ izražava vrijednost (*srednji arterijski tlak*). Pridjev *hormonski* ostvaruje čak četiri značenja: sastav (*hormonski preparat*), relaciju (*hormonski status*), uzrok (*hormonska promjena*) i funkciju (*hormonski receptor*), a pojavljuje se u tri vršne kategorije naziva: ⟨terapija⟩, ⟨simptom⟩ i ⟨ljudski organizam⟩.

- (7) Osobe koje imaju **infektivnu** mononukleozu imaju veći rizik od nastanka preosjetljivosti na ampicilin ili njegove analoge.
- (8) **Srednji** arterijski tlak je snižen.
- (9) Istraživanja na laboratorijskim životinjama usmjerena su i na ispitivanje djelovanja ED-a u tijeku fetalnog razvoja, koji nakon vezanja za **hormonske** receptore utječu na signalne putove i u konačnici dovode do mutacije DNA-a.

Prva hipoteza, da pridjevi u hrvatskome medicinskom nazivlju najčešće izriču vrijeme/trajanje, uzrok i lokaciju, djelomično je potvrđena. Prema korpusnim rezultatima najčešća su značenja pridjeva lokacija, svojstvo i relacija, a potom slijede uzrok, vrijeme/trajanje, funkcija i vrijednost. Najmanje su potvrđena značenja primjena, sastav, stanje, intenzitet, ishod i fizička značajka. Iako hipoteza o najčešćoj izraženosti vremena, uzroka i lokacije nije u potpunosti potvrđena, visoka frekvencija značenja lokacije, svojstva i relacije sugerira da pridjevi ponajprije služe preciznom određivanju prostorne i konceptualne orijentacije unutar medicinskoga znanja.



Grafikon 1. Najfrekventnija značenja pridjeva

Druga hipoteza, prema kojoj se značenje pridjeva u hrvatskome medicinskom nazivlju mijenja ovisno o kolokacijskome okruženju, nije potvrđena. Naime, čak 41 pridjev iz korpusa zadržava isto značenje u različitim kolokacijama. Manji broj pridjeva pokazuje značenjsku varijabilnost: sedam pridjeva ostvaruje dva značenja, dva pridjeva

(*negativni i primarni*) tri značenja, dok samo jedan pridjev (*hormonski*) ima četiri različita značenja u različitim kolokacijama.

Takav rezultat odstupa od početnih očekivanja s obzirom na teorijske pretpostavke o značenjskoj varijabilnosti pridjeva u stručnome diskursu. Jednoznačnost većine pridjeva može se tumačiti visokim stupnjem terminološke ustaljenosti i standardizacije u medicinskoj domeni, u kojoj kolokacijski obrasci odražavaju specijaliziranu i stabilnu upotrebu.

Treća hipoteza, da su najfrekventnije vršne kategorije naziva *bolest*, *terapija* i *simptom*, jest potvrđena. Među njima izrazito dominira kategorija *bolest*, što je i očekivano s obzirom na njezinu središnju ulogu u medicinskoj komunikaciji, dijagnostici i pojmovnoj klasifikaciji. Pridjevi u okviru te kategorije služe preciziranju, razgraničavanju i opisivanju različitih kliničkih stanja, što dodatno objašnjava njihovu učestalost.

S obzirom na ograničen opseg ovoga rada detaljniji prikaz rezultata dostupan je u Prilogu I, koji sadržava popis pridjeva prema kolokacijama, vršnim kategorijama naziva, primjerima iz korpusa i značenjima, te u Prilogu II, u kojemu su navedena značenja svih 50 pridjeva iz korpusa.

5. Zaključak

Analiza frekvencijskih obrazaca pridjeva u specijaliziranim korpusima omogućila je preciznije razumijevanje njihove funkcije u strukturiranju medicinskoga nazivlja te oblikovanju odnosa među stručnim pojmovima. Pridjevi se pritom ne pokazuju kao periferni jezični element, nego kao ključni nositelji terminološkoga značenja i klasifikacijskoga potencijala. Učestalost pojedinih pridjeva, njihova semantička razgranatost te pojava u višerječnim nazivima (npr. *akutna bol*, *kronična insuficijencija*, *maligni tumor*) odražavaju ne samo strukturna obilježja jezika struke nego i hijerarhiju značenja unutar medicinskoga znanja. Takve sveze upućuju na standardizirane opise i ustaljene obrasce, čime pridjevi aktivno doprinose procesu terminološke preciznosti, standardizacije i značenjske klasifikacije.

Korpusno utemeljena analiza kolokacija pokazala je da se značenja pridjeva ostvaruju u stvarnoj jezičnoj uporabi, a njihova varijabilnost i distribucija unutar vršnih kategorija naziva pridonose preciznijem opisu pojmova, većoj terminološkoj jasnoći te lakšem praćenju inovacija i trendova u medicinskome području.

Primjena korpusne metode pokazala se iznimno korisnom za analizu složenih obrazaca specijaliziranoga jezika. Od učitavanja tekstova, izrade konkordancija do identifikacije tipičnih kolokacija, ovaj pristup

omogućio je detaljnu analizu odnosa između pridjeva i imenica u medicinske nazivlje.

Ovo istraživanje pokazuje kako digitalne tehnologije, posebice korpusni alati, mogu poduprijeti razvoj i razumijevanje medicinskoga jezika. Praćenje učestalosti značenja pridjeva doprinosi standardizaciji hrvatskoga medicinskog nazivlja te osigurava preciznu, funkcionalnu i prilagodljivu komunikaciju u zdravstvenome kontekstu u skladu s novim znanstvenim spoznajama i potrebama struke.

Daljnja istraživanja mogla bi se usmjeriti na detaljniju značenjsku analizu većega broja pridjeva unutar pojedinih medicinskih potpodručja (npr. kardiologije, neurologije), čime bi se dodatno rasvijetlili obrasci značenjske specijalizacije i varijabilnosti. Također, moguće je istražiti značenjske promjene i inovacije u novijim nazivima s ciljem boljšeg razumijevanja dinamike terminološkoga razvoja. Dobiveni rezultati mogu poslužiti kao temelj za izradu terminoloških baza koje bi uključivale značenjsku razdiobu pridjeva i njihovu uporabu u različitim kontekstima, što bi bilo korisno i za leksikografske i za normativne svrhe.

Zahvalnica

Rad je izrađen u okviru istraživačkoga projekta Semantički okviri u hrvatskom jeziku (SEF), koji financira Europska unija - NextGeneration EU.

Literatura

- [1] Bourigault, D., & Slodzian, M. (1999). Pour une terminologie textuelle. *Terminologie Nouvelles*, 19, 29–32.
- [2] Bowker, L., Hawkins, S. (2006). Variation in the organization of medical terms: Exploring some motivations for term choice. *Terminology*, 12(1), 79–110.
- [3] Cabré, M. T. (1999). *Terminology: Theory, methods, and applications*. John Benjamins Publishing Company.
- [4] Cabré Castellvi, M. T. (2003). Theories of terminology: Their description, prescription and explanation. *Terminology*, 9(2), 163–199.
- [5] Cruse, A. (1986). *Lexical Semantics*. Cambridge University Press.
- [6] Dixon, R. M. W. (1982). *Where have all the adjectives gone? and other essays in semantics and syntax*. Mouton.
- [7] Dixon, R. M. W. & Aikhenvald, A. Y. (2004). *Adjective classes: A Cross-linguistic Typology*. Oxford University Press.
- [8] Durán-Muñoz, I. (2019). Adjectives and their keyness: a corpus-based analysis of tourism discourse in English. *Corpora*, 14(3), 351–378.

- [9] Faber Benítez, P. (2009). The cognitive shift in terminology and specialized translation. *MonTi. Monografías de Traducción e Interpretación, I*, 107–134.
- [10] Gil-Berrozpe, J. C.; León-Araúz, P. & Faber, P. (2019). Ontological Knowledge Enhancement in EcoLexicon, In I. Kosem et al. (Eds.), *Proceedings of the eLex 2019 conference* (pp. 177–197). Lexical Computing CZ.
- [11] Gjuran-Coha, A. (2011). Terminologizacija jezika medicinske struke. *Medicina Fluminensis*, 47(1), 4–14.
- [12] Gjuran-Coha, A. & Bosnar-Valković, B. (2013). Lingvistička analiza medicinskoga diskursa. *Jahr: europski časopis za bioetiku*, 4(1), 107–28.
- [13] Gjuran-Coha, A. & Bosnar-Valković, B. (2017). A study into the Croatian language of medicine. *English for Specific Purposes World*, 52, 1–18.
- [14] Grčić Simeunović, L. (2015). Prilog metodologiji opisa sintagmi u stručnom diskursu. *Rasprave: Časopis Instituta za hrvatski jezik i jezikoslovlje*, 41(1), 29–47.
- [15] Grčić Simeunović, L.; Stepišnik, U. & Vintar, Š. (2020). Klasifikacijska uloga pridjeva u domeni geomorfologije krša. *Rasprave: Časopis Instituta za hrvatski jezik i jezikoslovlje*, 46(2), 619–633.
- [16] Grčić, Simeunović, L. (2021). *Terminološki opis u službi stručnoga prevođenja. Dinamično modeliranje specijaliziranoga znanja*. Sveučilište u Zadru – Institut za hrvatski jezik i jezikoslovlje.
- [17] Kageura, K. & Umino, B. (1996). Methods of automatic term recognition. *Terminology*, 3(2), 259–289.
- [18] Kilgariff, A., Rychly, P., Kovar, V., Baisa, V. (2012). Finding multiwords of more than two words. In R. Vatvedt Fjeld & J. Matilde Torjusen (Eds.), *Proceedings of the 15th EURALEX International Congress* (pp. 693–700), Department of Linguistic and Scandinavian Studies.
- [19] L’Homme, M. C. (1998). Le statut de verbe en langue de spécialité et sa description lexicographique. *Cahiers de lexicologie* 73(2), 61–84.
- [20] L’Homme, M. C. (2003). Capturing the Lexical Structure in Special Subject Fields with Verbs and Verbal Derivates: A Model for Specialized Lexicography. *International Journal of Lexicography* 16(4), 403–422.
- [21] L’Homme, M. C. & Robichaud, B. (2014). Frames and terminology: Representing predicative terms in the field of the environment. In M. Zock, R. Rapp & C. R. Huang (Eds.), *Proceedings of the 4th Workshop on Cognitive Aspects of the Lexicon (CogALex)* (pp. 186–197), Association for Computational Linguistics and Dublin City University.

- [22] Lynch, C. (1997). Medical Terminology Management. In S. E. Wright & G. Budin (Eds.), *Handbook of Terminology Management I* (pp. 160–170), John Benjamins B.V.
- [23] Lyons, J. (1977). *Semantics*. Cambridge University Press.
- [24] Ljubešić, N. & Erjavec, T. (2011). hrWaC and slWac: Compiling Web Corpora for Croatian and Slovene. In I. Habernal, V. Matoušek (Eds.), *Text, Speech and Dialogue*, vol. 6836 (pp. 395–402), Springer Berlin Heidelberg.
- [25] Ljubešić, N. & Klubička, F. (2014). {bs,hr,sr} WaC – Web corpora of Bosnian, Croatian and Serbian. In F. Bidlhauer & R. Schäfer (Eds.), *Proceedings of the 9th Web as Corpus Workshop (WaC-9)* (pp. 29–35), Association for Computational Linguistics.
- [26] Pavić, M. (2022). *Uloga pridjeva u hrvatskome medicinskom nazivlju* [Doktorski rad, Sveučilište u Zagrebu]. Digitalni repozitorij Sveučilišta u Zagrebu. <https://repozitorij.ffzg.unizg.hr/islandora/object/ffzg%3A7165/datastream/PDF/view>
- [27] Pimentel, J. (2012). *Criteria for the Validation of Specialized Verb Equivalents: Application in Bilingual Terminography* [Doctoral dissertation: University of Montreal]. https://olst.ling.umontreal.ca/wp-content/uploads/olst/pdf/Pimentel_J_thesis_2012.pdf
- [28] Pitkänen-Heikkilä (2015). Adjectives as terms. *Terminology*, 21(1), 76–101.
- [29] Raffaelli, I. (2004). Polisemija pridjeva: sprega sintaktičkih i semantičkih struktura. *Filologija*, 44, 105–128.
- [30] Sager, J. C. (1990). *A practical course in terminology processing*. John Benjamins Publishing Company.
- [31] Taylor, B. R. (2017). *The Amazing Language of Medicine: Understanding Medical Terms and Their Backstories*. Springer International Publishing.
- [32] Temmerman, R. (2000). *Towards New Ways of Terminology Description. The Sociocognitive Approach*. John Benjamins Publishing Company.
- [33] Warburton, K. (2021). *The Corporate Terminologist*. John Benjamins Publishing Company.

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

Martina Pavić

Summary

This study analyzes the meanings of adjectives in Croatian medical terminology based on a specialized scientific corpus comprising 5.318,950 words, compiled and processed using the *Sketch Engine* tool. The primary aim of the research is to investigate the distribution, frequency, and meaning of adjectives in specialized collocations, as well as their function within the conceptual structure of medical knowledge.

The analysis focuses on 50 adjectives with over 100 occurrences each, including examples such as *acute*, *chronic*, *pathogenic*, *hormonal*, and *infectious*. Thirteen distinct meanings of adjectives were identified: location, relation, property, cause, time/duration, function, value, application, composition, condition, intensity, outcome, physical characteristic.

Adjectives were classified into three groups: those that change both the meaning and category of the term; those that shift the category but not the meaning; and those that preserve both. The majority of adjectives were found to be semantically astable across collocations, most commonly expressing the meanings of location, property, and relation, followed by cause and duration.

In terms of top-level conceptual domains, the most frequent were diseases, symptoms, and therapies, indicating their dominant role in structuring medical discourse. These findings supports the view that adjectives are not merely peripheral modifiers, but rather integral elements of terminological precision and semantic differentiation. Their role in classifying and specifying core concepts reflects their importance in the hierarchical and paradigmatic organization of specialized knowledge.

Furhermore, the corpus-based collocational approach proves valuable for identifying regularities in lexical behaviour, which may inform both terminological standardization nad future lexicographic or didatic appllications in medical language.

Key words: adjective meanings, Croatian medical terminology, scientific corpus, collocations, top-level conceptual domains

Prilog I. Tablice pridjeva prema kategorijama naziva, kolokacijama i značenju

A) Pridjevi koji ne mijenjaju kategoriju naziva ni značenje

pridjev	kategorija naziva	kolokacije	primjer iz korpusa	značenje
patogeni	mikroorganizam	mikroorganizmi, bakterije	Većina ovakvih pneumonija nastaje aspiracijom patogenih mikroorganizama iz orofaringealnog sekreta.	UZROK
peroralni	terapija	pripravak, kortikosteroid, glukokortikoid	Ako se bolesnikovo stanje poboljša, treba utvrditi najnižu dnevnu dozu peroralnih glukokortikoida uz koju će astma biti dobro kontrolirana.	PRIMJENA
masni	dio tijela	jetra, tkivo, stanica	Ono može utjecati na razvoj inzulinske neosjetljivosti, nastanak metaboličkog sindroma, a lučenjem adipokina, čimbenika stvorenih u masnim stanicama, doprinosi upalnim procesima, pa se i iz tog razloga pretilost smatra kroničnom upalnom bolešću.	SASTAV

B) Pridjevi koji mijenjaju kategoriju naziva, ne i značenje

pridjev	kategorija naziva	kolokacije	primjer iz korpusa	značenje
alergijski	bolest	rinitis, dermatitis, konjunktivitis, upala, sindrom	Najčešći oblik je cjelogodišnji alergijski rinitis, često praćen očnim simptomima.	UZROK
	nuspojava	reakcija	Kod jedne bolesnice zbog opsežne urtikarije i moguće alergijske reakcije na lijek odlučeno je primijeniti lijek u sljedećoj dozi uz individualnu procjenu doze kliničkog farmakologa.	
invazivni	bolest	karcinom, trihosporonoza, adenokarcinom	U većini opisanih slučajeva invazivne trihosporonoze (63%), u podlozi su bile maligne bolesti krvi kao glavni predisponirajući činitelj.	SVOJSTVO
	postupak	mehanička ventilacija, zahvat, metoda	Prosječna duljina trajanja invazivne mehaničke ventilacije iznosila je 14,7 sati.	
intravenski	terapija	imunoglobulin	Osim antibiotika, postoje istraživanja koja su pokazala korist od davanja intravenskih imunoglobulina (IVIG) u slučaju pneumokokne septe.	PRIMJENA
	postupak	urografija, injekcija	U slučaju recidiva stenoze radili smo ponovno intravensku urografiju.	
	medicinska oprema	kontrast	Snimke se rade prije i nakon uporabe intravenskog kontrasta ovisno o postavljenoj sumnji koji je organ ozlijeđen u arterijskoj, venskoj i/ ili odgođenoj fazi.	

pridjev	kategorija naziva	kolokacije	primjer iz korpusa	značenje
kronični	bolest	rinitis, kolecistitis, gastritis	Patohistološkim nalazom želučane sluznice utvrđen je umjereni kronični gastritis s blagom žarišnom atrofijom. Druge moguće nuspojave su kronični umor, osip, svrbež te hiperuricemija.	VRIJEME/ TRAJANJE
	nuspojava	umor		
periferni	bolest	neuropatija, tumor, artritis	U razdoblju djetinjstva i adolescencije često se zbog periferne neuropatije javlja simptom akroparestezija koji se prezentira akutnom ili kroničnom boli u šakama i stopalima.	LOKACIJA
	dio tijela	krv, neuroni, živčani sustav	U perifernom živčanom sustavu serotonin ima vazokonstriktorski i vazodilatacijski učinak i sudjeluje u normalnom ciklusu seksualnog odgovora kroz seksualno uzbuđenje.	
poslijeoperacijski	postupak	praćenje, oporavak	Naime, izvrstan rezultat postignut je u 25-ero od 27-ero (92,5%) liječenih bolesnika, a izvijestili su o samo jednom recidivu (3,7%), uz napomenu da je prosječno poslijeoperacijsko praćenje bilo 67 mjeseci (između 24 i 128 mjeseci).	VRIJEME/ TRAJANJE
	bolest	poremećaj, komplikacija	Jedina zabilježena poslijeoperacijska komplikacija uključuje produženo krvarenje.	

pridjev	kategorija naziva	kolokacije	primjer iz korpusa	značenje
pozitivni	dijagnoza	nalaz, obiteljska anamneza	Pozitivni nalazi na BKV nađeni su u 6 (17 %) bolesnika, dok su svi testovi na JCV bili negativni.	ISHOD
	postupak	hemokultura	Prema različitim istraživanjima pozitivna hemokultura bilježi se od 5 do 33 % pacijenata, češće u djece nego u odraslih.	
	ljudski organizam	protutijela	Smatra se općenito da bolesnici s pozitivnim protutijelima imaju lošiju prognozu od bolesnika u kojih nisu dokazana protutijela.	
	mikroorganizam	izolati	Svi pozitivni izolati testiraju se na INH, RIF, STR, EMB i PZA.	
terapijski	postupak	pristup, intervencija	Međutim, kombinacijom spomenutih terapijskih pristupa postiže se bolji učinak i značajno se smanjuje mortalitet i morbiditet bolesnika.	RELACIJA
	terapija	odgovor, učinak, doza	Bolesnike za interferonsku terapiju treba odabrati prema vjerojatnosti terapijskog odgovora.	

C) Pridjevi koji mijenjaju kategoriju naziva i značenje

pridjev	kategorija naziva	kolokacije	primjer iz korpusa	značenje
gnojni	bolest	meningitis, peritonitis	Također, ukoliko sumnjamo da bolesnik ima gnojni meningitis treba procijeniti je li potrebna prethodna slikovna obrada (kompjuterizirana tomografija ili magnetska rezonancija mozga) prije lumbalne punkcije.	SVOJSTVO
	simptom	sekrecija, iskašljaj	U bolesnika s konjuktivitisom vodeći simptomi bili su injekcija spojnice (94 %), suze-nje očiju (57 %), gnojna sekre-cija (48 %) i fotofobija (33 %).	SASTAV
hormonski	terapija	kontracepcija, preparat	Naglašava se i povećana učestalost EN-a povezana s promjenom 370 hormonskog statusa u žena kao posljedica uporabe hormonske kontracepcije, odnosno tijekom trudnoće ili dojenja.	SASTAV
	simptom	status	Osim hormonskog statusa glavni prijeoperacijski simptom bila je glavobolja kod više od trećine (trideset šest bolesnika) i kod vrlo malog broja (osamnaest bolesnika) vidne smetnje.	RELACIJA
		promjena	Kod žena su posljedice još izraženije u menopauzi, kada se hormonske promjene odražavaju u obliku osteoporoze koja uzrokuje smanjivanje visine trupova kralježaka.	UZROK
	ljudski organizam	receptor	Istraživanja na laboratorijskim životinjama usmjerena su i na ispitivanje djelovanja ED-a u tijeku fetalnog razvoja, koji nakon vezanja za hormonske receptore utječu na signalne putove i u konačnici dovode do mutacije DNA-a.	FUNKCIJA

pridjev	kategorija naziva	kolokacije	primjer iz korpusa	značenje
infektivni	bolest	mononukleoza, endokarditis, gangrena	Osobe koje imaju infektivnu mononukleozu imaju veći rizik od nastanka preosjetljivosti na ampicilin ili njegove analoge.	UZROK
	mikroorganizam	serovar	Analizirajući rezultate dobivene serološkim pretraživanjem ljudi i životinja tijekom 2007. godine može se zamijetiti veća učestalost vjerojatnih infektivnih serovara Australis i Grippotyphosa.	SVOJSTVO
velik	bolest	nepovoljna događanja	Velika nepovoljna događanja (MACE) definirana kao razvoj srčanog infarkta i/ili revaskularizacija ciljne lezije i/ili smrt pojavili su se u 1,6% bolesnika.	INTENZITET
	dio tijela	mozak	Da bi došlo do razvoja osjeta boli potrebno je da bolni podražaj od mjesta ozljede preko perifernih živaca i kralježnične moždine stigne u centar za bol u velikom mozgu (thalamus).	FIZIČKA ZNAČAJKA

Prilog II. Prikaz pridjeva i njihovih značenja u hrvatskome medicinskom nazivlju

A) Pridjevi s jednim značenjem

pridjev	značenje	kolokacija
akutni	VRIJEME/ TRAJANJE	bronhitis, infarkt miokarda, pankreatitis, hepatitis, bol
alergijski	UZROK	rinitis, dermatitis, konjunktivitis, reakcija, sindrom, upala
blagi	INTENZITET	rinitis, hepatitis, infekcija
degenerativni	SVOJSTVO	artritis, promjena, bolesti
desni	LOKACIJA	klijetka, ventrikul, atrij
dijagnostički	FUNKCIJA	kriterij, postupak, obrada, pretraga, metoda, paracenteza, protokol
intraoperacijski	VRIJEME/ TRAJANJE	ventilacija, monitoring, autotransfuzija, tromboliza, dijaskopija, torakotomija
intravenski	PRIMJENA	imunoglobulin, aplikacija, infuzija, urografija, ubrizgavanje, injekcija
invazivni	SVOJSTVO	karcinom, tumor, trihosponoza, adenokarcinom, metoda, mehanička ventilacija, tehnika, zahvat
izvanbolnički	LOKACIJA	pneumonija, srčani zastoj, srčani arest, pacijent, hitna medicinska služba, fibrinoliza, reanimacija
kirurški	RELACIJA	zahvat, postupak, bolesnik, odjel
koronarni	LOKACIJA	arterije, bolest, ateroskleroza, cirkulacija, protok
kronični	VRIJEME/ TRAJANJE	rinitis,olecistitis, hepatitis, prostatitis, gastritis, umor, kašalj
krvni	RELACIJA	slika, tlak, grupa, žila, ugrušak
laboratorijski	RELACIJA	pretraga, dijagnostika, pokazatelj, test, analiza, medicina, biljeg
lijevi	LOKACIJA	klijetka, ventrikul, atrij
limfni	RELACIJA	čvor, žile, drenaža
lokalni	LOKACIJA	anestetik, infekcija, režanj, ishemija, liječenje
mali	FIZIČKA ZNAČAJKA	krvne žile, mozak, tumor, arterije
maligni	SVOJSTVO	melanom, tumor, bol, stanica, proces
masni	SASTAV	jetra, tkivo, stanica
niski	VRIJEDNOST	krvni tlak, temperatura, mortalitet, patogenost, infektivnost, viremija
patogeni	UZROK	mikroorganizmi, bakterije, hantavirus, rikecije

pridjev	značenje	kolokacija
periferni	LOKACIJA	neuropatija, arteriopatija, artritis, tumor, krv, živac, arterije, vene, puls, matične stanice
perkutani	PRIMJENA	drenaža, nefrostomija, termorizotomija
peroralni	PRIMJENA	kortikosteroid, antihistaminik, glukokortikoid
pozitivni	ISHOD	nalaz, obiteljska anamneza, hemokultura, titar
probavni	RELACIJA	sustav, cijev, trakt, smetnja
rezistentni	SVOJSTVO	mikroorganizam, bakterija, hipertenzija, kvazispesijes, tuberkuloza, pneumokok
terapijski	RELACIJA	pristup, odgovor, intervencija, učinak, doza
unutarnji	LOKACIJA	uho, organ
upalni	STANJE	odgovor, proces, parametar, citokin, prostatitis
venski	LOKACIJA	sustav, staza, put, hipertenzija, insuficijencija, tlak, centralni kateter
virusni	UZROK	hepatitis, pneumonija, infekcija, gastroenteritis
visoki	VRIJEDNOST	krvni tlak, temperatura, mortalitet, patogenost, letalitet, infektivnost

B) Pridjevi s dva značenja

antimikrobni	FUNKCIJA SVOJSTVO	terapija, liječenje, sredstvo, lijek, profilaksa, djelovanje rezistencija, osjetljivost
arterijski	LOKACIJA RELACIJA	tlak, kateter, krv hipertenzija
bubrežni	LOKACIJA RELACIJA	transplantat, presađivanje, transplantacija, cista, tubul, tkivo, čašica bolesnik, funkcija, disfunkcija
gnojni	SVOJSTVO SASTAV	meningitis, peritonitis iscjedak, iskašljaj, rana
infektivni	UZROK SVOJSTVO	bolesti, mononukleoza, endokarditis, gangrena serovar
ishemijski	SVOJSTVO UZROK	moždani udar, kardiomiopatija, bol oštećenje
kardiovaskularni	LOKACIJA FUNKCIJA	bolest, sustav, kolaps, anomalija, zahvat, poremećaj lijek
srednji	LOKACIJA VRIJEDNOST	uho arterijski tlak
veliki	INTENZITET FIZIČKA ZNAČAJKA	nepovoljna događanja krvne žile, trohanter, arterije, mozak, tumorska masa

C) Pridjevi s tri značenja

negativni	VRIJEDNOST STANJE ISHOD	tlak, test, hemokultura simptom anamneza, test
primarni	FUNKCIJA VRIJEME/TRAJANJE STANJE	prevencija, radioterapija, rezistencija, tumor limfedem, glavobolja hipertenzija

D) Pridjevi s četiri značenja

hormonski	SASTAV SVOJSTVO RELACIJA UZROK	preparat kontracepcija, liječenje, terapija status, receptor disbalans, promjena

Quality Textual Corpora and New South Slavic Language Models

Scientific paper

DOI: 10.18485/judig.2025.1.ch19

Mihailo Škorić¹,  0000-0003-4811-8692

Saša Petalinkar²,  0009-0007-9664-3594

Abstract

This paper presents a newly published quality corpus, *Znanje*, which builds on the existing S.T.A.R.S. corpus of Serbian academic research, but extends its range to more South Slavic languages, primarily Croatian and Slovenian. The new corpus is compiled using texts from the S.T.A.R.S corpus, as well as Croatian academic writings collected from the DABAR repository, and Slovenian academic writings collected from several online repositories under the Open Science Slovenia initiative. Secondly, the paper presents two new encoding language models for Serbo-Croatian, trained using this dataset, as well as the *Kišobran* umbrella web-corpus of Serbian and Croatian, and Wikipedia and Wiki source dumps for Serbo-Croatian, Serbian, Croatian and Bosnian. Finally, it discusses the evaluation of these new models on several previously established text vectorization tasks (for Serbian and Serbo-Croatian), their comparison with the similar models previously trained in this language domain and the outlook of preparing large language models for South Slavic languages in the future.

Key words: language models, corpora, Serbian, Croatian, parameter-efficient training.

¹ Language Resources and Technologies Society (JeRTeh), mihailo@jerteh.rs

² Language Resources and Technologies Society (JeRTeh), sasa5@jerteh.rs

1. Introduction

The performance and reliability of language models are fundamentally tied to the quality of the textual data they are trained on. High-quality corpora ensure that models learn accurate linguistic patterns, semantic relationships, and contextual nuances, which are essential for tasks such as text classification, translation, summarization, and semantic search. In low-resource languages or dialects, such as South Slavic languages, the scarcity of corpora in general, but also especially well-curated corpora, has historically limited the development of robust models.

Recent research underscores that not only the size but also the diversity, cleanliness, and representativeness of the corpus significantly affect model generalization and downstream task performance. Poorly curated datasets can introduce noise, bias, and inconsistencies, leading to models that misinterpret or misrepresent linguistic phenomena (*Yang et al., 2025*). For multilingual models, the imbalance in corpus quality across languages can further exacerbate bias and reduce cross-lingual transfer efficiency (*Xu et al., 2024*).

Therefore, constructing a high-quality academic corpus is a critical step toward advancing language technologies for South Slavic languages. It not only addresses the data scarcity issue but also lays the groundwork for training fairer and more effective language models in this linguistic domain.

This research builds upon (*Škorić & Janković, 2024*), where several new corpora for Serbian and Serbo-Croatian were introduced, including a doctoral dissertation corpus, S.T.A.R.S. and a parallel academic abstract corpus, *PaSaž*. Doctoral dissertations represent a highly valuable resource for building academic corpora due to their rigorous standards in methodology, linguistic quality, and peer review. These texts offer a rich and diverse thematic range across scientific disciplines and are often timely in addressing contemporary research questions. In Serbia, the NaRDuS³ repository provides open access to thousands of doctoral theses, supported by legal mandates and standardized metadata. This makes it an ideal foundation for constructing a large, high-quality scientific corpus in Serbian.

To compile the corpus, metadata for all dissertations was programmatically retrieved and enriched with additional fields to assess document accessibility, language, OCR requirements, and copyright status. Dissertations were filtered to include only those written in Serbian, available in full-text PDF format, not requiring OCR, and not under restrictive copyright. This process yielded 11,624 dissertations or approximately 87.5% of the repository. The final corpus extracted using

³ <https://nardus.mpn.gov.rs>

PyMuPDF⁴ library and cleaned to retain only paragraph text, comprised over 560 million words, making it one of the largest and most reliable academic corpora in Serbian.

Since the initial compilation of the doctoral dissertation corpus, S.T.A.R.S has been further expanded by incorporating research papers (peer-reviewed journal articles, conference proceedings, and other scholarly publications) from multiple academic online repositories in Serbia, i.e. active institutional repositories of higher education institutions, where researchers and peers store the published papers (*Otašević, 2023*). This expansion increased the size of the corpus to over 700 million words.

That work established the methodological framework for corpus compilation, which this paper extends to a broader South Slavic context by incorporating Croatian and Slovenian academic texts.

2. New Corpus

To further enrich the corpus and extend its linguistic and thematic coverage, additional academic texts were gathered from repositories across several countries of former Yugoslavia. For each country, prominent open-access academic repositories were identified, and custom scraping scripts were developed to extract relevant documents. Beyond the full-text PDFs, the collection process prioritized key metadata fields essential for corpus structuring and future filtering: title, creator, source, subject (where available), PDF URL, and language. In cases where language metadata was missing, particularly in smaller repositories, manual annotation was performed to ensure consistency and usability across the dataset.

For the Croatian portion of the corpus, academic texts were sourced from DABAR (Digital Academic Archives and Repositories)⁵, the national infrastructure for institutional repositories in Croatia. DABAR hosts a wide range of scholarly outputs, including theses, dissertations, and research papers from Croatian universities and research institutions. Its standardized metadata and open-access structure made it an ideal candidate for large-scale corpus construction. Despite this, DABAR had not previously offered a publicly available, structured corpus suitable for language modelling. Using a custom scraping pipeline tailored to DABAR's architecture, a total of 108,786 documents were collected, comprising approximately 53.4 million sentences and over 1.2 billion words. This substantial addition significantly enhances the corpus's linguistic diversity and supports the development of more robust language models for Croatian.

⁴ <https://pymupdf.io>

⁵ <https://dabar.srce.hr>

For Bosnia and Herzegovina, the corpus was expanded using academic texts from a single well-structured and accessible repository: the institutional repository of the University of East Sarajevo⁶. This repository provided a consistent metadata schema and open access to a range of academic documents, making it suitable for automated collection. A tailored scraping script was developed to extract both full-text PDFs and essential metadata. In total, 8 documents were collected, containing 21,156 sentences and approximately 510,000 words. While modest in scale compared to other national segments, this addition contributes valuable linguistic and thematic diversity to the overall corpus, as the texts were written in Ijekavian Serbian. The repository of the University of Zenica⁷ was also considered, but eventually the effort was scrapped, since these texts were mostly written in English, and metadata was scarce and .

For Montenegro, academic texts were collected from the institutional repository of the University of Montenegro⁸, the country's only public university. This repository provided a centralized and structured source of scholarly documents, making it suitable for targeted scraping. A custom script was developed to extract both full-text PDFs and essential metadata: title, author, source, and language. In total, 315 documents were collected, comprising approximately 605,119 sentences and 14.5 million words. This contribution marks the first large-scale academic corpus from Montenegro and adds valuable linguistic representation to the South Slavic dataset.

For Slovenia, academic repositories were defined through the Open Science Slovenia initiative⁹, which aggregates institutional repositories from major Slovenian universities and research centers. The repositories of the University of Ljubljana¹⁰, University of Maribor¹¹, University of Primorska¹², and University of Nova Gorica¹³ were included, along with specialized repositories such as DiRROS (Digital Repository of Research Outputs in Slovenia)¹⁴ and REVIS (Repository of Educational and Scientific Works)¹⁵. These platforms provided structured metadata and open access to a wide range of academic documents. It should be noted that unlike other South Slavic countries, Slovenia had already previously

6 <https://repozitorijum.ues.rs.ba>

7 <https://www.epub.unze.ba>

8 <https://eteze.ucg.ac.me>

9 <https://www.openscience.si>

10 <https://repozitorij.uni-lj.si>

11 <https://dk.um.si>

12 <https://repozitorij.upr.si>

13 <https://repozitorij.ung.si>

14 <https://dirros.openscience.si>

15 <https://revis.openscience.si>

established a large-scale academic corpus from Open Science Slovenia, *OSS 1.0*, which reports over 3.2 billion words from more than 150,000 scientific texts (Erjavec, Fišer & Ljubešić, 2021). During this effort, however, for the Serbo-Croatian portion, 176 documents were collected, totaling 113,339 sentences and 2.7 million words. For Slovenian-language texts, a significantly larger dataset was compiled: 148,158 documents, comprising 104.9 million sentences and over 2.29 billion words. This makes the Slovenian segment the largest national contribution to the corpus and a major resource for future Slovenian language modeling efforts.

For North Macedonia, no suitable academic repository was identified that met the criteria for inclusion in the corpus. The institutional repositories of Goce Delčev University¹⁶ and St. Clement of Ohrid University¹⁷ were reviewed, but both presented challenges in terms of metadata accessibility and document structure. Manual inspection revealed that only a small number of documents were written in South Slavic languages, with many texts either lacking language metadata or being authored in English, Albanian and Greek. Due to these limitations, Macedonia was not included in the current version of the corpus, though future expansions may revisit this if repository infrastructure improves.

Table 1 presents contributions to the Serbo-Croatian portion of the final corpus by repository source or sources. The total word count of this portion is nearly 2 billion, making it the largest compiled academic corpora for these languages. This, however, still falls short of the Slovenian portion, which has a word count of nearly 2.3 billion.

Table 1: Total contributions to the Znanje corpus from each source repository (Serbo-Croatian macro-language portion that was used for model training).

Source	Doc. count	Sent. count	Word count	Share
NARDUS	11,432	22,779,252	574,600,000	30%
Institutional repositories in Serbia	10,889	4,192,656	109,400,000	5.70%
University of Montenegro	315	605,119	14,500,000	0.80%
University of East Sarajevo	8	21,156	510,000	< 0.1%
DABAR	108,786	53,369,657	1,214,000,000	63.40%
Open Science Slovenia	176	113,339	2,694,331	0.10%
Total	131,606	81,060,023	1,915,704,331	100%

¹⁶ <https://eprints.ugd.edu.mk>

¹⁷ <https://repository.ukim.mk>

3. New Models

Building on the Znanje corpus¹⁸, we trained two new transformer-based language models using the *XLM-R* architecture (Conneau et al., 2020). These models are specifically designed for the Serbo-Croatian macro-language, and thus only Serbo-Croatian portion of the Znanje corpus was used. To further enrich the training data, we also incorporated the umbrella web corpus Kišobran¹⁹, which encompasses all published web-crawled datasets of Serbian, Croatian, Bosnian and Montenegrin texts, as well as compiled Wikipedia dumps for Serbian²⁰, Croatian, and Serbo-Croatian, and Wikisource dumps for Serbian, Croatian, and Bosnian²¹. These sources provided a diverse and balanced representation of contemporary and formal language use across the macro-language spectrum. In addition to these modern sources, the training data included the SrpELTeC corpus²², a collection of digitized Serbian literary texts from the 19th and early 20th centuries (*Stanković, Krstev, Todorović & Škorić, 2021*). This corpus, consisting of historical novels and prose, added valuable diachronic depth to the language model and helped capture stylistic and lexical variation over time. Altogether, the combined training dataset exceeded 20 billion words, making it the largest corpus ever used for training South Slavic language models. The resulting models were published on Hugging Face²³, ensuring open access for further research and application in NLP tasks involving South Slavic languages.

To preserve the multilingual capabilities of the original *XLM-R* architecture while minimizing computational costs, we adopted parameter-efficient fine-tuning using LoRA (Low-Rank Adaptation) methods (*Hu et al., 2021*). LoRA enables selective updating of a small subset of model parameters, allowing the model to adapt to new data without full retraining. This approach was particularly suitable given the scale of the training corpus, which was over 20 billion words, and the goal of maintaining compatibility with the broader multilingual *XLM-R* framework. By using LoRA, we achieved efficient specialization for Serbo-Croatian while retraining the model's generalization capacity across related languages.

To accommodate different use cases and computational environments, two model variants were developed: *XMLi*²⁴ and

¹⁸ <https://huggingface.co/datasets/procesaur/znanje>

¹⁹ <https://huggingface.co/datasets/procesaur/kisobran>

²⁰ <https://huggingface.co/datasets/procesaur/Vikipedija>

²¹ <https://huggingface.co/datasets/procesaur/Vikizvornik>

²² <https://huggingface.co/datasets/jerteh/SrpELTeC>

²³ <https://huggingface.co>

²⁴ <https://huggingface.co/te-sla/XMLi>

*TeslaXLM*²⁵. *XLMini* is a compact model based on the *XLMini-R-base* architecture, with 279 million parameters, optimized for faster inference and lower resource consumption, ideal for real-time applications and deployment on limited hardware. *TeslaXLM*, on the other hand, is built on the *XLMini-R-large* architecture, with 561 million parameters, offering superior performance on complex NLP tasks due to its greater representational capacity. This dual-model approach ensures flexibility, allowing researchers and developers to choose between speed and precision depending on their needs. Due to usage of LoRA, models were trained updating only 0.03% of the model parameters, significantly reducing computational demands. The training was conducted on a setup of four NVIDIA A4000 GPUs, with *XLMini* completing in 52 hours and *TeslaXLM* in 186 hours.

4. Evaluation and Results

To assess the performance of the newly trained models (*XLMini* and *TeslaXLM*) alongside three existing baselines: *XLMini-R-base* and *XLMini-R-large* (Conneau et al., 2020), as well as *XLMini-R-BERTić* (Ljubešić et al., 2024). We conducted evaluations on two previously defined upstream tasks (Škorić, 2024). The first task was masked token prediction in literary texts: *The Adolescent* by Fyodor Mikhailovich Dostoevsky represented by two different Serbian translations²⁶, and *Around the World in 80 Days* by Jules Verne in both Serbian and Croatian²⁷. All models use the same tokenizer, which was also employed to mask the input texts, ensuring consistency across evaluations. This task tested the models' ability to recover contextually appropriate words in literary prose, a challenging domain due to its stylistic and lexical complexity.

The second task focused on sentence pair recognition, evaluating semantic similarity and alignment. This was performed on parallelized versions of the same novels: *The Adolescent* for Serbian–Serbian pair recognition, and *Around the World in 80 Days* for Serbian–Croatian pair recognition. These evaluations measured how well each model could distinguish between semantically aligned and misaligned sentence pairs while also reflecting their capacity for cross-lingual understanding within the Serbo-Croatian macro-language. The results of both tasks are presented in Figures 1 and 2, showing comparative performance across all five models.

²⁵ <https://huggingface.co/te-sla/TeslaXLM>

²⁶ Mladić, Beograd (1972), Dečko, Beograd (1988)

²⁷ Put oko sveta za 80 dana, Beograd (1962), Put oko svijeta u osamdeset dana, Zagreb (1961)

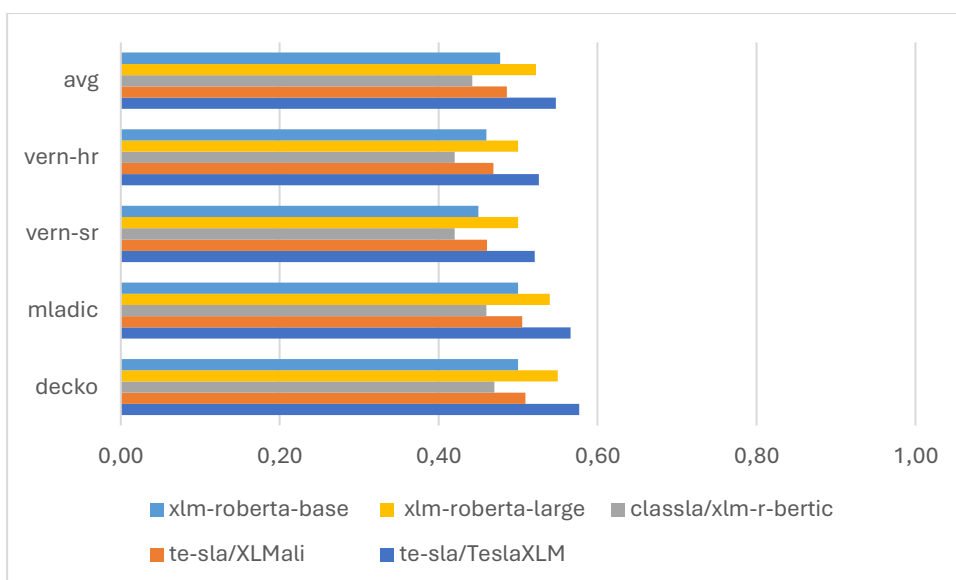


Figure 1: Masked token prediction accuracy across five models on literary texts in Serbian and Croatian. Evaluation performed on masked excerpts from Dostoevsky’s *The Adolescent* and Verne’s *Around the World in 80 Days*.

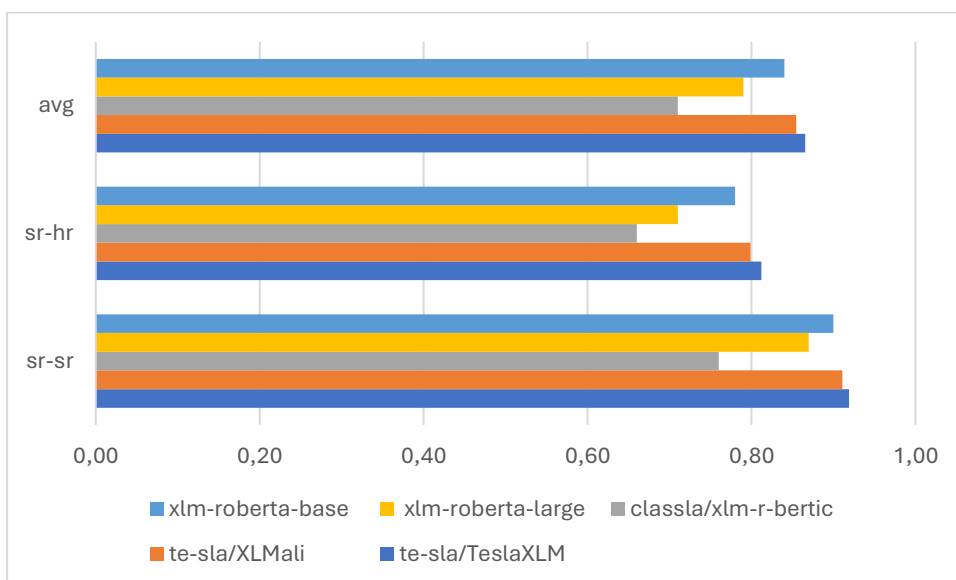


Figure 2: Sentence pair recognition performance for Serbian–Serbian and Serbian–Croatian parallel texts. Models evaluated on an aligned and misaligned sentence pairs from *The Adolescent* and *Around the World in 80 Days*, measuring semantic similarity and cross-lingual understanding.

On the first task *TeslaXLM* performed best in all categories, likely due to its larger parameter count and domain-specific training. *XLM-R-large* also performed well, showing that size and multilingual pretraining help. *XLMali*, while smaller, still outperformed *XLM-R-BERTić* and *XLM-R-base*, validating its efficiency-focused design. On the second task, *TeslaXLM* again leads, showing strong cross-lingual semantic understanding. Interestingly, *XLMali* slightly outperforms *XLM-R-base* and *large*, suggesting that targeted fine-tuning on Serbo-Croatian data yields better results than general multilingual pretraining.

In addition to upstream evaluations, all five models were tested on a suite of five downstream NLP tasks to assess their practical utility and generalization capabilities. These tasks included part-of-speech (POS) tagging (17 classes) using SrpKor4Tagging dataset²⁸, named entity recognition (NER, 5 classes) using SrpELTeC-gold-NER²⁹, sentiment classification (3 classes) (*Stanković et al., 2022*), as well as emotion classification (8 classes), and moral classification (5 classes)³⁰. Each task was selected to reflect a different aspect of linguistic understanding, from syntactic structure to affective and ethical interpretation.

The results of these evaluations are presented in Figure 3, which compares model performance across all five tasks. These benchmarks provide insight into how well each model adapts to real-world applications in South Slavic language processing. Notably, *TeslaXLM* consistently outperformed other models in tasks requiring deeper semantic understanding such as emotion and moral classification, while in the other categories results were similar. *XLMali* also offered competitive results with significantly faster inference times, but on average fell shortly out of *XLM-R-base* and *XLM-R-Bertić* on downstream tasks.

5. Conclusion

This research presents a significant advancement in the development of language resources and models for South Slavic languages. By compiling the *Znanje* corpus, which integrates academic texts from Serbia, Croatia, Bosnia and Herzegovina, Montenegro, and Slovenia, we have created one of the most comprehensive and linguistically diverse corpora in the region. The inclusion of over 20 billion words from curated sources such as doctoral dissertations, national repositories, Wikipedia, and historical literature marks a milestone in corpus construction for low-

²⁸ <https://huggingface.co/datasets/jerteh/SrpKor4Tagging>

²⁹ <https://huggingface.co/datasets/jerteh/SrpELTeC-gold-NER>

³⁰ Unpublished annotated datasets

resource languages. This effort not only fills a critical gap in available training data but also lays the groundwork for future multilingual and cross-lingual NLP research in the South Slavic domain.

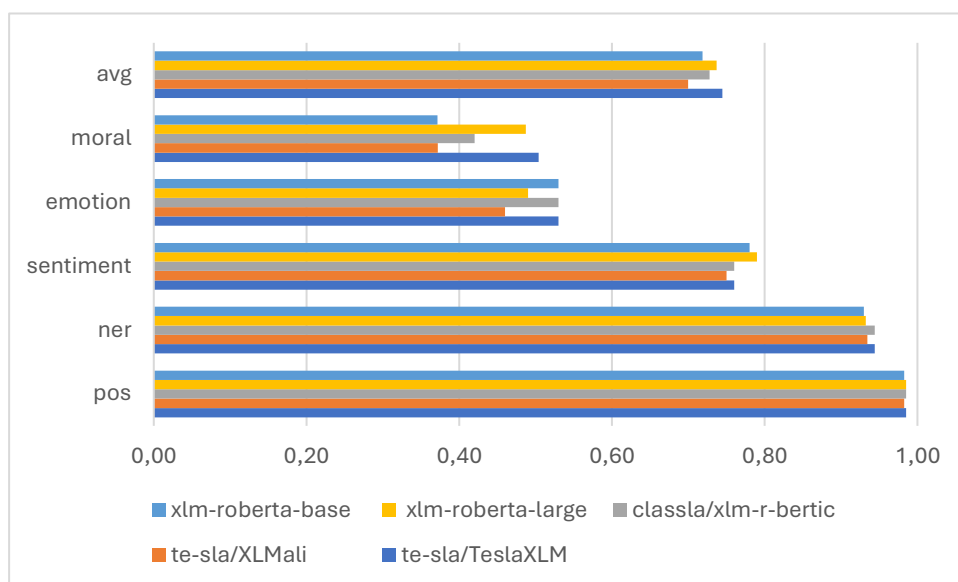


Figure 3: Downstream task performance comparison across five models. Tasks include part-of-speech tagging (17 classes), named entity recognition (5 classes), sentiment classification (3 classes), emotion classification (8 classes), and moral classification (5 classes).

The development of two new Serbo-Croatian language models, XLMali and TeslaXLM, demonstrates the effectiveness of parameter-efficient fine-tuning using LoRA, enabling high performance with reduced computational cost. Evaluation across upstream and downstream tasks shows that these models outperform or match existing multilingual baselines, particularly in semantic understanding and classification tasks. By publishing the models on Hugging Face, we ensure open access and reproducibility, supporting further research and practical applications.

While the Znanje corpus and the newly trained models represent a major step forward for South Slavic NLP, several avenues remain open for future research. Expanding the corpus to include underrepresented languages such as Macedonian and minority dialects within the Serbo-Croatian continuum would further enhance linguistic coverage.

Acknowledgment

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

References

- [1] Xu, Y., Hu, L., Zhao, J., Qiu, Z., Ye, Y., & Gu, H. (2024). *A Survey on Multilingual Large Language Models: Corpora, Alignment, and Bias*. arXiv. <https://arxiv.org/pdf/2404.00929>
- [2] Yang, S., Jing, M., Wang, S., Kou, J., Shi, M., Xing, W., Hu, Y., & Zhu, Z. (2025). *Exploring Large Language Models in Healthcare: Insights into Corpora Sources, Customization Strategies, and Evaluation Metrics*. arXiv. <https://arxiv.org/abs/2502.11861>
- [3] Škorić, M., & Janković, N. (2024). *New Textual Corpora for Serbian Language Modeling*. Infotheca. - Vol. 24, No. 1 (2024), p. 71–96.
- [4] Otašević, V. *Transfer of Metadata into the National Information System of Scientific Research Activities with Automatic Authorship Association*. Infotheca. - Vol. 23, No. 2 (2024), p. 27–48.
- [5] Erjavec, T., Fišer, D., & Ljubešić, N. (2021). *The KAS corpus of Slovenian academic writing*. Language Resources and Evaluation, 55(2), 551-583.
- [6] Conneau, A. et al. (2019). *Unsupervised cross-lingual representation learning at scale*. arXiv preprint arXiv:1911.02116.
- [7] Stanković, R., Krstev, C., Todorović, B. Š., & Škorić, M. (2021). *Annotation of the serbian eltec collection*. Infotheca–Journal for Digital Humanities, 21(2), 43-59.
- [8] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2022). *Lora: Low-rank adaptation of large language models*. ICLR, 1(2), 3.
- [9] Ljubešić, N., Suchomel, V., Rupnik, P., Kuzman, T., & van Noord, R. (2024). *Language Models on a Diet: Cost-Efficient Development of Encoders for Closely-Related Languages via Additional Pretraining*. arXiv preprint arXiv:2404.05428.
- [10] Škorić, M. (2024). *New Language Models for Serbian*. Infotheca. - Vol. 24, No. 1 (2024), p. 7–28.
- [11] Stanković, R., Košprdić, M., Nešić, M. I., & Radović, T. (2022, June). Sentiment analysis of Serbian old novels. In *Proceedings of the 2nd Workshop on Sentiment Analysis and Linguistic Linked Data* (pp. 31-38).

Квалитетни текстуални корпуси и нови јужнословенски језички модели

Михаило Шкорић, Саша Петалинкар

Овај рад представља значајан допринос развоју језичких ресурса и модела за јужнословенске језике. Описује конструисање новообјављеног корпуса *Знање*, који обухвата академске текстове из Србије, Хрватске, Босне и Херцеговине, Црне Горе и Словеније. Корпус је изграђен комбиновањем више извора: докторских дисертација из НаРДУС-а, научних радова из националних репозиторијума као што су ДАБАР (Хрватска) и *Open Science Slovenia*, као и институционалних репозиторијума широм Србије, универзитета у Источном Сарајеву и Универзитета Црне Горе. Корпус се може грубо поделити на део који чине текстови на српскохрватском макро-језику (нешто мање од две милијарде речи) и део на словеначком језику (нешто мање од 2,3 милијарде речи).

На основу српскохрватском дела овог корпуса, и неколико других (Кишобран веб-корпус, Википедија, Викизворник и СрпЕЛТеК) обучена су два нова језичка модела за векторизацију текста: *XLMAli* (279 милиона параметара) и *TeslaXLM* (561 милион параметара), као адаптације *XLM-R* архитектуре. Модели су обучени су уз помоћ LoRA (Low-Rank Adaptation) технике и циљањем свега 0.3% параметара, што је омогућило ефикасну обуку уз минималне рачунарске ресурсе. Модели су тестирани на низу задатака: попуњавање маскираних речи у књижевним текстовима, препознавање парова реченица, као и пет задатака обраде природног језика — обележавање врстом речи, препознавање именованих ентитета, класификација сентимената, емоција и моралних вредности.

Резултати показују да *TeslaXLM* постиже најбоље резултате у већини задатака, док *XLMAli* нуди одличан баланс између брзине и прецизности. Упоредна анализа са постојећим моделима као што су *XLM-R-base*, *XLM-R-large* и *XLM-R BERTi* потврђује предности доменски специфичне обуке. Модели су објављени на платформи *HuggingFace* и доступни су за даљу употребу и истраживање. Рад отвара простор за будући развој језичких технологија за мање заступљене језике и дијалекте, као и за проширење корпуса и модела на македонски и друге варијетете.

Кључне речи: Језички модели, корпуси, српски језик, хрватски језик, параметарски ефикасно учење.

Sintaksičko-semantička anotacija elektronskog korpusa srpskog jezika

Naučni rad

DOI: 10.18485/judig.2025.1.ch20

Nataša Kiš¹,  0000-0001-8627-0316

Apstrakt

U radu će biti predstavljene neke od mogućnosti korišćenja elektronskog korpusa kao primarnog izvora građe za različita istraživanja iz oblasti sintakse i semantike savremenog srpskog jezika. Dva osnovna zadatka rada jesu da se prikažu dosadašnja istraživanja vezana za proces komplementacije prideva i pridevskih imenica, odnosno njihovog rekcijskog potencijala na sintagmatskom planu, a koja počivaju na relevantnoj građi preuzetoj prvenstveno iz elektronskog *Korpusa savremenog srpskog jezika*, SrpKor2013 (<https://noske.jerteh.rs/#dashboard?corpname=SrpKor2013>). Drugi zadatak je da se ukaže na potrebu proširenja mogućnosti pretrage elektronskih korpusa vezanu za sintaksičko-semantičku analizu na nivou proste rečenice. Pretrage ovog tipa zasnivaju se prvenstveno na sintaksičkoj i semantičkoj anotaciji korpusa i uključivale bi, između ostalog, podatke o valencijskim osobinama predikata rečenice i svim obaveznim i fakultativnim aktantima u konkretnoj situaciji, a važan segment jeste izdvajanje semantičkih klasa reči i njihovih sintaksičko-semantičkih osobenosti na osnovu kojih bi se mogli pretraživati različiti rečenični modeli. Jedan od ciljeva rada jeste da se ukaže na odgovarajuće lingvističke analize koje bi omogućile izdvajanje i uključivanje podataka potrebnih za dalji razvoj u oblasti sintaksičke i semantičke anotacije korpusa.

Ključne reči: sintaksa, semantika, anotacija, elektronski korpus, rekcija, srpski jezik

¹ Odeak za srpski jezik i lingvistiku, Filozofski fakultet, Univerzitet u Novom Sadu, natasab@ff.uns.ac.rs

1. Uvod

1. U opisu svakog pojedinačnog prirodnog jezika, bilo da se radi o deskriptivnoj odnosno sinhronijskoj lingvistici, istorijskoj odnosno dijahronijskoj lingvistici, komparativnoj ili pak kontrastivnoj lingvistici, od izuzetnog je značaja uvid u što obimniju i stilski što raznovrsniju jezičku građu. Razvoj korpusne lingvistike umnogome doprinosi relevantnosti istraživanja i utemeljenosti zaključaka analiza u svim oblastima lingvističkih istraživanja. Ovom prilikom ukazaćemo na mogućnosti korišćenja elektronskog korpusa kao primarnog izvora građe za određena istraživanja iz oblasti sintakse i semantike savremenog srpskog jezika.

Dva osnovna zadatka rada jesu da se prikažu dosadašnja istraživanja vezana za proces komplementacije prideva i pridevskih imenica, odnosno njihovog rekcijaskog² potencijala na sintagmatskom planu, a koja počivaju na relevantnoj građi preuzetoj prvenstveno iz elektronskog *Korpusa savremenog srpskog jezika*, SrpKor2013 (<https://noske.jerteh.rs/#dashboard?corpname=SrpKor2013>). Drugi zadatak je da se ukaže na potrebu proširenja mogućnosti pretrage korpusa zasnovanu na sintaksičkoj i semantičkoj anotaciji. Pretrage ovog tipa omogućavale bi pristup podacima o valencijskim karakteristikama predikata rečenice i svim obaveznim i fakultativnim aktantima u konkretnoj situaciji, rekcijским особеностима глагола, prideva, imenica i priloga, a važan segment takve analize može biti izdvajanje semantičkih klasa reči i njihovih sintaksičko-semantičkih особености на основу којих би се могли преtraživati i različiti rečenični modeli.

2. Primena korpusa SrpKor2013 u lingvističkim istraživanjima

2. U studiji *Dopune prideva u savremenom srpskom jeziku* (Kiš 2016) posmatran je odnos komplementacije u pridevskoj sintagmi sa primarnim ciljem da se ukaže na osnovne sintaksičko-semantičke karakteristike odnosa komplementacije prideva u savremenom srpskom jeziku. Istraživanje je podrazumevalo da se najpre izdvoje oni pridevi koji dobijaju dopune prema inherentnoj leksičkoj semantici, u svim ili samo u nekim od svojih značenja, kako bi se mogla izvršiti semantička kategorizacija ovih prideva. Stoga je bilo je potrebno odabrati recentni rečnik srpskog jezika, mi smo se opredelili za šestotomni rečnik Matice

² U domaćoj i stranoj literaturi ima različitih određenja pojma *rekcija*. Mi ćemo ovaj odnos posmatrati kao „посебан, граматикализован, облик зависног односа који се првенствено сагледава на синтагматском нивоу, између две пунозначне (аутосемантичне) речи. Једна граматички доминантна реч – глагол, именица, придев – без обзира на то у ком је облику употребљена, управља датом везом, а друга је регирана” (Ружић у: Пипер и др., 2005: 537).

srpske (RMS), i na osnovu značenja prideva izdvojili one za koje se moglo očekivati da zahtevaju dopunu. Sledeći korak bio je pronalaženje svih izdvojenih prideva upotrebljenih u različitim jezičkim kontekstima. Pokazalo se da je svakako najadekvatniji izvor građe upravo elektronski *Korpus savremenog srpskog jezika*, SrpKor2013, dostupan na Matematičkom fakultetu Univerziteta u Beogradu,³ budući da je omogućavao pretragu u tekstovima koji pripadaju svim funkcionalnim stilovima savremenog jezika, a što je omogućilo praćenje aktuelnih zbivanja i tendencija u savremenom srpskom jeziku, posebno kada se uzmu u obzir i oni primeri koji nisu u skladu sa jezičkom normom. Pri tom, prilikom formiranja upita postoji mogućnost da se pretražuju samo izvorni tekstovi, kao i da budu uključeni i prevodi. Analizirano je blizu 600 prideva, dok korpus samog rada sačinjava 383 prideva,⁴ a iz elektronskog korpusa je ekscerpirano gotovo 9000 rečenica.

2.1. Prilikom pretrage korpusa bilo je potrebno izdvajati čitavu paradigmu jednog prideva, kao i sve rečenice u kojima se data lema u korpusu pojavljivala budući da je u analizi odnosa komplementacije relevantan prvenstveno jezički kontekst.⁵ Regularni izraz u pretrazi predstavlja gramatičku osnovu⁶ prideva, međutim kako bi bili izdvojeni i oblici prideva u nominativu muškog roda sa nepostojanim A, primenjivan je zasebni upit: **sigurn[a-z]*** i **siguran**. Korpus omogućava više načina pretraživanja ukoliko se koriste operatori grupisanja, npr. ako se žele pretražiti svi oblici flektivne paradigme imenice deca, ponuđena su dva načina: **deca|dece|deci|decu|deco|decom** ili **dec(a|e|i|u|o|om)**. Međutim, kako je paradigma prideva veoma razučena zbog kongruencije sa imenicama u tri gramatička roda i dva broja, opredelili smo se za navedeni oblik upita iako je njegov nedostatak što su na taj način izdvajane i druge leme sa istim korenom, npr. *sigurnost*, *sigurnosni* i sl. Nakon uvida u moguće realizacije tražene pridevske lekseme i sintagmatske spojeve čiji je upravni ili glavni član pridev, pretraga je bila sužena tako što su desno od prideva dodavani ili predlozi ili pak prototipični veznik dopunske klauze

³ U svojoj doktorskoj disertaciji Miloš Utvić (2014: 247–303) detaljno predstavlja nastanak i razvoj elektronskog *Korpusa savremenog srpskog jezika*. Autor izdvaja parametre za verziju SrpKor2013 i pokazuje faze u kreiranju ovog korpusa: prikupljanje tekstova i njihovu digitalizaciju, obradu elektronskih tekstova, izvršene anotacije – bibliografska, strukturna i morfološka, indeksiranje tekstova i pripremu paralelizovanih korpusa.

⁴ Za realizaciju dopuna uz neke od prideva, uprkos njihovoj inherentnoj leksičkoj semantici, koja za to ima potencijala, i na osnovu koje su pridevi inicijalno izdvajani, nisu pronađene potvrde u korpusu, pa tako nisu bili ni obuhvaćeni analizom.

⁵ Uticaj nejezičkog konteksta svakako je važan, ali sam elektronski korpus ne pruža uvid u njega.

⁶ Gramatička osnova prideva dobija se kada se pridevu u genitivu jednine, bez obzira na rod, ukloni nastavak za genitiv (sigurn-og, sigurn-e, sigurn-og).

da. Sami predlozi bili su deo očekivane predložko-padežne konstrukcije u funkciji dopune prideva: **sigurn[a-z]* u; sigurn[a-z]* da**.

Iste večeri pevala sam sa ranjenicima , [sigurnija u](#) sebe , kao da sam raskinula sa nečim što je ležalo duboko u meni .

Prvo, kao što će biti objašnjeno u narednom poglavlju , mi još nismo sasvim [sigurni u](#) to koja tačno teorija uspešno objedinjuje opštu relativnost i kvantnu mehaniku,

Da li ste [sigurni da](#) želite da SAJLAS izvrši ovaj zadatak?

Na ovaj način iz pretrage bi bili isključeni komplementi u slobodnoj padežnoj formi – bez predloga, ili blokirani obaveznim determinatorom, a koji bi mogli biti izdvajani samo prilikom šire pretrage koja bi podrazumevala prvobitni regularni izraz upita.⁷ Sintaksička i semantička anotacija⁸ korpusa mogle bi unaprediti i pojednostaviti pretrage potrebne za različite vrste analiza na ovom jezičkom nivou, o čemu će biti više reči u narednom odeljku.

Nakon formiranja korpusa za istraživanje zadatak je bio opisivanje sintaksičkih i semantičkih osobina samih prideva radi utvrđivanja semantičkih i sintaksičkih tipova dopuna koji se pojavljuju u funkciji komplemenata pojedinih semantičkih klasa prideva.

2.1.2. Već ova faza istraživanja pokazala je potrebu izdvajanja *tvorbenih gnezda/porodica* kako bi se mogle pratiti sintaksičko-semantičke karakteristike korenske/motivne morfeme i motivisanih reči. Mogućnost pretrage u elektronskim korpusima preko formiranih tvorbenih gnezda omogućila bi različite vrste analiza i na drugim jezičkim nivoima. U tom smislu, kada se radi o morfološkoj anotaciji⁹ korpusa, od značaja bi bilo grupisanje lema koje predstavljaju prvo-stepene, drugo-stepene ... derivate izvedene iz jedne proste, korenske morfeme. Kao problem u pretragama tog tipa mogu se pojaviti *alomorfi* iste morfeme,¹⁰ na primer:

⁷ Sam pridev *siguran* ne regira dopune u vidu slobodnih padežnih formi, kao što je to slučaj sa pridevima tipa *pun, ravan, siromašan, svestan* i sl.

⁸ Pod pojmom anotacije korpusa podrazumeva se „postupak kojim se dodatne informacije pridružuju delovima korpusa (tekstovi, logičke celine u okviru teksta, tokeni) (Utvić 2014: 27). Prema vrstama podataka razlikuje se više nivoa anotacije, a ovom prilikom su u fokusu *sintaksička anotacija* (pridruživanje jednočlanom ili višečlanom nizu korpusnih reči zajedničke informacije o funkciji u rečenici (subjekat, predikat, objekat, glagolska odredba) ili obeležavanje takvog niza kao sintagme (imenička, pridevska, itd.)) i *semantička anotacija* (pridruživanje oznaka odgovarajućeg značenja svakom tokenu) (Utvić 2014: 28).

⁹ Miloš Utvić predstavio je u svom magistarskom radu osnovne modele morfološke obrade elektronskog teksta (2008: 22–25).

¹⁰ Na ovaj segment morfološke realizacije leksema skreće pažnju i Miloš Utvić (2008: 20–21).

ZVAti – ZOVnuti – ZIVkati; PLOT – PLETivo – prePLITati;
REĆi – REKoh – REĆi – REČ; VUĆi – VUKoh – VUČem – VUCi;
NOGa – NOŽi – NOŽni

2.1.3. Još jedan važan parametar prilikom analize procesa komplementacije i rekcijaskog potencijala prideva jeste rečenična funkcija u kojoj se realizuju. Naime, pridev može da dobije dopunu ukoliko se nalazi u poziciji leksičkog jezgra predikata. Stoga su analizom obuhvaćeni pridevi koji se pre svega nalaze u leksičkom jezgru kopulativnog i semikopulativnog predikata, ali i pridevi koji u rečenici imaju funkciju apozitiva ili predikativnog atributa, a koji su inače kondenzatori rečeničnog sadržaja. Navešćemo prideve upotrebljene u ovim pozicijama u rečenici, a primerima ćemo pokazati njihovu sposobnost vezivanja dopuna:

- Pridev u kopulativno-imenskom predikatu (Popularnu kulturu stvaraju podređeni koji *nisu pomirljivi* sa svojim položajem u društvu)
- Pridev u semikopulativnom predikatu (Kao *merodavne* za dokazivanje vlasništva, Udruženje smatra gruntovnice iz 1939. godine)
- Pridev kao apozitiv (Aprilska penzija, skromna i *nedovoljna* za najosnovnije potrebe, stigla je i već je, kažu, potrošena)
- Pridev kao predikativni atribut (Vojnici je s ljubavlju primaju, *zahvalni* na njenoj pažnji)

2.2. Govoreći o parsiranju, Miloš Utvić navodi da se ovaj proces u užem smislu „poistovećuje sa (automatskom) sintaksičkom analizom pomoću računara ‘čiji je zadatak prepoznavanje rečenice i pridruživanje (odgovarajuće) sintaksičke strukture’” (navedeno prema Utvić 2014: 106 [Jurafsky & Martin, 2008: 427]).

Parsiranje je bitan zadatak u automatskoj obradi teksta na prirodnom jeziku i često je neophodan preduslov za naknadne obrade i primene poput semantičke analize, analize diskursa, ekstrakcije informacija, mašinskog prevođenja, itd. Takođe, parsiranje je složen zadatak i postoji više različitih pristupa rešavanju tog problema, okupljenih u klase sintaksičkih teorija, među kojima su najznačajnije *gramatike frazne strukture* ili *konstituentne gramatike* (eng. phrase structure grammars, constituency grammars) i *gramatike zavisnosti* (eng. Dependency grammars)) (Utvić 2014: 106).

Kada se radi o analizi odnosa komplementacije na sintagmatskom nivou, naša dosadašnja istraživanja upućuju na zaključak da je parsiranje zasnovano na gramatici zavisnosti adekvatnija teorijska osnova.

Rezultat sintaksičke analize gramatikom zavisnosti se takođe predstavlja drvetom zavisnosti, pri čemu:

- čvorovi drveta, izuzev korena, su morfosintaksičke reči u rečenici;
- (usmerene) grane drveta predstavljaju zavisnosti i usmerene su ka zavisnoj reči,
- svaka grana je obeležena odgovarajućom gramatičkom funkcijom ili semantičkom relacijom;
- koren drveta je specijalan čvor koji je granom povezan sa glagolom kao centralnom („glavnom”) rečju u rečenici (Utvić 2014: 108).

U slučaju gramatika zavisnosti ([Mel’čuk, 1988]) nad skupom morfosintaksičkih reči u okviru rečenice uspostavlja se binarna relacija takva da je morfosintaksička reč *x* u relaciji sa morfosintaksičkom rečju *y* ako i samo ako *y* „zavisi” od *x*, pri čemu svaka od zavisnosti predstavlja neku gramatičku funkciju (na primer subjekat, objekat, itd.) ili opštu semantičku relaciju (agent, cilj, itd.) (Utvić 2014: 108).

Prethodna precizna lingvistička analiza procesa komplementacije, kao i drugih sintaksičko-semantičnih procesa i relacija u prosto rečenici neophodan je uslov za preciznu sintaksičku anotaciju elektronskih korpusa.¹¹ Time bi se, na primer, prilikom određivanja svake od zavisnosti, odnosno svake gramatičke funkcije, mogla praviti razlika između njihovih podtipova: subjekat (gramatički i logički), objekat (direktni i indirektni – tipični i netipični) i dopune (obavezne, fakultativne i obavezni determinatori).

Istraživanja sintaksičkih i semantičkih karakteristika deadjektivnih odnosno pridevskih imenica trenutno su u toku. Jedan od važnih segmenata ovog istraživanja, koji se velikim delom odnosi na rekcijski potencijal deadjektivnih imenica, ima za cilj takođe izradu rekcijskog rečnika. Analizom je obuhvaćena 361 imenica (broj imenica je manji od broja prideva koji regiraju dopune jer se, primera radi, od nekih prideva ne mogu derivirati deadjektivne imenice (*drugačiji*, *znan*, *izmirljiv*, *isti*, *kadar*, *konten*, *ravan*, *saodnosan*), a sama pretraga elektronskog korpusa otežana je činjenicom da se rekcijski potencijal deadjektivnih imenica razlikuje u manjoj ili većoj meri od rekcijskog potencijala motivnih prideva. Za to postoje brojni, prvenstveno jezički faktori, ali se ovom prilikom na tome nećemo zadržavati.

¹¹ Osnova za sintaksičko parsiranje može se tražiti, između ostalog, u rekcijskim rečnicima koji su urađeni za veliki broj slovenskih jezika. Radi se mahom o rekcijskim rečnicima glagola, ali i imenica i prideva. Na konferenciji Leksikografija i leksikologija u svetlu aktuelnih problema održanoj 2020. godine u Beogradu, u organizaciji Instituta za srpski jezik SANU, autorka je predstavila model budućeg rekcijskog rečnika prideva (Киш 2021). Pregled ove vrste specijalnih rečnika u slovenskim lingvističkim tradicijama videti u: Киш 2021; Штрбац 2010.

Osnovni upit predstavlja regularni izraz koji čini paradigma jedne leme – to su u ovom slučaju svi padežni oblici deadjektivne imenice u jednini i množini. Upit čini gramatička osnova deadjektivne imenice koja se dobija odbijanjem padežnog nastavka u genitivu jednine: **bezobzirnost[a-z]***. Na ovaj način dobija se uvid u moguće sintaksičke, odnosno gramatičke funkcije ovih imenica u rečenici, inventar potencijalnih predložko-padežnih formi dopuna, uvid u dopune realizovane bez predloga, a moguće je pratiti u kojem se jezičkom kontekstu date imenice realizuju sa dopunama ali i bez njih. Precizna sintaksička anotacija korpusa mogla bi olakšati ovakve vrste pretraga ukoliko bi, na primer, upit omogućavao da se izdvajaju samo imenice prema rečeničnim funkcijama ili pak prema sintagmama čiji su upravni, glavni član.¹² Kako je moguće pretragu suziti samo na slučajeve kada deadjektivne imenice vezuju uz sebe dopune sa strukturnim likom B (imenica blokirana predlogom),¹³ regularni izraz formiran je na isti način kao prilikom pretrage prideva: npr. **bezobzirnost[a-z]* prema**.

Nasilnička vožnja među najtežim prekršajima Nasilnička vožnja, kao jedan od najtežih prekršaja prema novom zakonu, predstavlja nepoštovanje pravila i [bezobzirnost prema](#) ostalim učesnicima u saobraćaju.

Pri formiranju upita mora se uzeti u obzir i činjenica da, za razliku od prideva, dopune deadjektivnih imenica ne moraju biti u kontaktnom položaju sa imenicama (često se imenice realizuju sa različitim determinatorima, npr. determinator kojim se iskazuje nosilac osobine, stanja ili raspoloženja i sl.), tako da se podrazumeva drugačiji regularni izraz: **velikodusnost[a-z]* [{0,6} prema**.

To bi vrlo lepo istaklo nesebičnost i [velikodusnost pređašnjih očeva i uopšte pređašnjih ljudi prema](#) novoj, lakomislenoj socijalnoj omladini.

Ono što je za pretragu elektronskog korpusa relevantno jeste i mogućnost izvođenja više deadjektivnih imenica, najčešće dve, od istog

¹² Prilikom sintaksičke anotacije korpusa bilo bi korisno, na primer, uzeti u obzir dekomponovane i perifrastične predikate, koji bi se mogli posmatrati kao spojevi u kojima se ostvaruje različit stepen kolokacijske veze između semikopulativnih glagola i deverbativnih i deadjektivnih imenica u njihovom sastavu (npr. *držati govor, izneti predlog, nemati vrednost, uliti hrabrost* i sl.) (Радовановић / Бугарски, 2005: 206–207). O kolokacijama u korpusnoj lingvistici videti u: Utvić 2014: 195–198).

¹³ O strukturi padežnog sistema srpskog jezika videti detaljnije u: Ivić 1995: 195–203.

prideva. Ove imenice razlikuju se u manjoj ili većoj meri prema svojim semantičkim i sintaksičkim karakteristikama, a neretko jedan od oblika nije zabeležen u RMS, što je svakako predmet naše analize. Ovakvi slučajevi su, čini se, još jedna potvrda potrebe da se u elektronskom korpusu pretrage rade prema tvorbenim gnezdim/porodicama. Samo jedan od brojnih primera jesu imenice *adaptivnost* (ovaj oblik nije zabeležen u RMS) i *adaptacija*:

Bez obzira na diktirane uslove , neophodno je imati u vidu koliko duh naše tradicije ima realne [adaptivnosti](#) za recepciju onog što se zvanično prihvata.

Sredstva ove aproprijacije namenjena su za rekonstrukciju i [adaptaciju](#) verskih objekata na teritoriji Republike Srbije , obuhvatajući AP Kosovo i Metohija

... donošenje sektorskih planova i programa mera za [adaptaciju](#) na klimatske promene i usaglašavanje sektorskih strategija sa strategijama , belom knjigom EU za adaptaciju i relevantnim direktivama EU ;

2.3.1. Naredni primer dovodi nas do još jednog problema koji bi mogao dobiti svoje razrešenje u sintaksički i semantički anotiranom elektronskom korpusu. Od prideva *vredan* izvedene su dve imenice – *vrednost* i *vrednoća*.

vredan, -dna, -dno

1. marljiv, radan.
2. koji vredi, koji ima vrednost, dragocen.
3. a. koji zasluđuje (nešto), dostojan (čega) .
b. kadar, sposoban.

vrednost

1. svojstvo onoga što je vredno, što je na ceni.
2. ono što vredi.
3. važnost, značaj.
4. valjanost, punovažnost.
5. veličina, znatnost; dar, sposobnost.
6. mat. količina.

vrednoća

- osobina onoga koji je vredan, marljivost, radinost, valjanost.

Imenica *vrednoća* može se dovesti u semantičku vezu sa pridevima koji imaju značenje *marljiv*, *radan*, i poput prideva, dobija dopunu kojom se data osobina vezuje samo za određene situacije, domene života neke osobe u kojima se ispoljava. Značenja imenice *vrednost* motivisana su drugim značenjskim likom istog prideva, a sve se to ogleda na sintaksičko-semantičkom planu i u rekcijским konstrukcijama koje grade ove lekseme.

Ovo je rezultat vredan *uvažavanja*.

Te dve knjige su vredne da se pročitaju.

Kupili su akcije vredne 14 miliona dolara.

Ovo zapažanje je *od velike* vrednosti za istoričare

Svoj visoki položaj duguje svojoj vrednoći u učenju.

Kao relevantan parametar za razlikovanje ovih imenica pojavljuje se semantička kategorija *animatnosti* (živo+/-) pojma nosioca osobine. Naime, nosilac osobine označene imenicom *vrednoća* markiran je obeležjem živo+, dok s druge strane, za pojam kojem se pripisuje neka *vrednost* ovo obeležje nije relevantno.

3. Mogućnosti pretrage u sintaksički i semantički anotiranim korpusima

3. Ovi primeri vode nas do drugog zadatka ovog rada. Namera nam je da se iz ugla analize na nivou proste rečenice ukaže na potrebu proširenja mogućnosti pretrage anotiranih korpusa.

Pretraga korpusa bila bi svakako preciznija ukoliko bi se pri sintaksičkoj anotaciji uzeli u obzir podaci o valencijskim osobinama¹⁴ predikata rečenice i svim obavezanim i fakultativnim aktantima u konkretnoj situaciji, pri čemu bi trebalo uključiti i različite, relevantne sintaksičko-semantičke parametre kao što su: *animatnost*, *usmerenost/direktivnost*, *recipročnost*, *ablativnost*, *obuhvatnost*, *partitivnost* i sl. Na taj način, mogle bi biti izdvojene grupe prideva koji su markirani, primera radi, obeležjem *recipročnosti* i koji, upotrebljeni kao leksičko jezgro različitih vrsta analitičkog predikata, utiču na sintaksičku strukturu rečenice i semantičke relacije u njoj: A) pridevi **jednakosti** – *ekvivalentan*, *identičan*, *istovetan*, *jednak*, *kongruentan*, *homologan*, *korespondentan*, *podudaran*, *proporcionalan*, *ravan*, *ravnopravan*, *recipročan*, *saglasan*, *sazvučan*, *saobrazan*, *srodan*, *konkurentan*, *isti*, *izjednačen*, *konkordantan*, *konforman*, *sinoniman*, *ujednačen*, *uporediv*, *usklađen*, *adekvatan*; B) pridevi **nejednakosti**, **različitosti**

¹⁴ Valentnost u užem, jezičkom smislu, može se odrediti kao spajanje, povezivanje ili kombinovanje glagolskih i imenskih reči i izraza na sintaksičko-semantičkom planu. Ovakvo udruživanje reči zasnovano je na leksičosemantičkoj vezi među njima, ali i na gramatičkim pravilima koja važe za svaki jezik posebno. Broj učesnika u situaciji odnosno broj aktanata zavisi prvenstveno od valencije predikata (Ружић у: Пипер и др., 2005: 517).

– *kontradiktoran, nesličan, protivan, protivrečan, obrnut, suprotan, nesaglasan, nesrodan, oprečan, kontraran, različit, drugačiji, nepodudaran*; C) pridevi **prostornih odnosa** – *naporedan, paralelan, simetričan, pripijen, priljubljen*; D) pridevi sa značenjem **socijativnosti** – *dobar, intiman, nerazdvojan, nesložan, blizak, zaručen, inkompatibilan, komplementaran, korelativan, uzdržan, nepovezan, nespojiv, nevezan, neuporediv, suplementan, kompatibilan, saodnosan*.

Kada je pridev *jednak* deo imenskog predikata, u rečenici će biti dva obavezna aktanta između kojih se uspostavlja relacija iskazana pridevom – nosilac osobine i drugi nosilac osobine. Oni mogu biti različito raspoređeni u rečenici, odnosno mogu biti u tačno utvrđenim funkcijama u rečenici, pri tom sintaksičko-semantičko obeležje *animatnost* (živo+/-) utiče na mogućnost uspostavljanja relacije između pojmova sa obeležjem živo+ ili živo-:

- oba učesnika u poziciji gramatičkog subjekta (*Ana i Ivana su jednake, Trougao A i trougao B su jednaki*) ili je prvi učesnik u poziciji gramatičkog subjekta, a drugi je u poziciji dopune sa značenjem koučesnika u situaciji (*Ana je jednaka Ivani / sa Ivanom; Trougao A jednak je trouglu B / sa trouglom B*).

Relacioni pridevi *paralelan* i *zaručen* pojavljuju se u rečenicama strukturiranim prema istom modelu, ali pridev *paralelan* podrazumeva da učesnici u relaciji imaju obeležje živo-, dok pridev *zaručen* uspostavlja odnos među učesnicima sa oznakom živo+:

- *Duž A i duž B su paralelne; Duž A paralelna je duži B / sa duži B*
- *Marko i Ivana su zaručeni; Marko je zaručen za Ivanu / sa Ivanom*.

3.1. Ako pogledamo pridev *veran* i jedan od primarnih derivata, imenicu *vernost*, videćemo razlike u njihovom rekcijском potencijalu. Relacioni pridev *veran* vezuje uz sebe objekatsku dopunu u dativu bez predloga sa oznakom živo+/- (on je *veran prijatelju / toj ideologiji*), imenica *vernost* dobija isti tip objekatske dopune realizovane različitim formama (pokazao je *vernost prijatelju / prema prijatelju*), ali se može u određenom kontekstu upotrebiti i bez dopune (oni ne mogu da govore o *vernosti*).

3.2. Posebnu pažnju treba skrenuti na višeznačnost leksema koja se odražava i na rekcijски potencijal reči budući da jedna leksema u svojim različitim semantičkim realizacijama može imati i različita sintaksičko-semantička obeležja što može biti problem kako u semantičkoj tako i u sintaksičkoj anotaciji korpusa.¹⁵ Pridev *veran* i imenica *vernost*, ukoliko

¹⁵ Višeznačnost i obrada „nep(rep)oznatih” reči jesu glavni problemi u automatskoj morfosintaksičkoj anotaciji korpusa, pri čemu se u ovom kontekstu može reći da „višeznačnost (eng. ambiguity) nastupa kada se jednom tokenu može pridružiti više morfosintaksičkih opisa koji se međusobno isključuju” (Utvić 2014: 143).

nosilac osobine ima oznaku živo-, mogu imati značenje „biti jednak nečemu“, što implicira da će dopuna koju regiraju pripadati drugačijem semantičkom tipu, odnosno imaće značenje drugog nosioca osobine (tekst je *veran originalu*; izdavač se držao *vernosti prema tekstu*).

3.3. Iz ugla govornika srpskog kao maternjeg ili kao nematernjeg/stranog jezika bilo bi korisno da se u pretragama elektronskog korpusa može poći od semantičkih klasa reči čiju bi formalizaciju, odnosno strukturne modele u kojima se realizuju pokazali primeri ekscerpirani iz elektronskog korpusa (npr. u klasi reči sa značenjem optativne modalnosti nalaze se pridev *željan*, imenica *želja* i glagol *želeti*, koji se mogu realizovati u sledećim rečeničnim modelima: Ja sam *željna odmor* – Moja *želja za odmorom* je velika / Ja imam *želju za odmorom* – Ja *želim da se odmorim / odmor*).

3.4. Posebno važno i svakako kompleksno pitanje jeste način formalizacije različitih semantičkih uloga u rečenici, o čemu govori Miloš Utvić predstavljajući proces parsiranja zasnovanog na gramatici zavisnosti (2014: 108–111). Tako je, na primer, u strukturiranju rečenice, relevantno i pitanje kojim se jezičkim sredstvima iskazuju podaci o agensu/pseudoagensu. U savremenom srpskom jeziku izdvajaju se različiti strukturni tipovi s obzirom na ove parametre, tako da, ako u rečenici postoji doživljavač/eksperiment, rečenica može biti strukturirana na više načina: *ja bih jela jabuke* / *ja želim da jedem jabuke* / *jedu mi se jabuke* i sl.).

4. Zaključak

Iz svega rečenog, može se videti koliko su sintaksička i semantička anotacija elektronskog korpusa izuzetno kompleksni procesi koji zahtevaju postojanje teorijskih osnova analize vezanih za svaki pojedinačni prirodni jezik. Gramatike pojedinačnih jezika, čak i ako se radi o istim jezičkim porodicama, na sintaksičko-semantičkom planu često se značajno razlikuju. To svakako ima uticaja na način obrade i parsiranja tekstova. Stoga je, s jedne strane, saradnja između lingvista i onih koji se bave programerskim poslovima neophodna, kako bi se ukazalo na potrebne jezičke analize koje bi omogućile i dale osnov za efikasniju i precizniju anotaciju korpusa. S druge strane, mogućnosti primene sintaksički i semantički anotiranih korpusa u različitim jezičkim istraživanjima, ali i u nastavi jezika, ne samo kao maternjeg, već i kao nematernjeg, odnosno stranog, u značajnoj bi meri unapredile i učinile efikasnijim istraživačke i nastavne procese.

Izvori

- [1] *Речник српскохрватског књижевног језика* (1967–1976). Нови Сад / Загреб: Матица српска / Матица хрватска (Матица српска – Матица хрватска 1-3, Матица српска 4-6).
- [2] Vitas, D. i Utvić, M. *Korpus savremenog srpskog jezika* (verzija SrpKor2013) Copyright Grupa za jezičke tehnologije Univerziteta u Beogradu. <https://noske.jerteh.rs/#dashboard?corpname=SrpKor2013>

Literatura

- [1] Ivić, M. (1995). *Lingvistički ogledi*. Biblioteka XX vek, knj. 61. Slovoğraf.
- [2] Jurafsky, D. & Martin, J. H. [2008]. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*, second ed., Prentice Hall.
- [3] Киш, Н. (2016). *Допуне придева у савременом српском језику*. Едиција Е-дисертација, књ. 11. Филозофски факултет. <http://digitalna.ff.uns.ac.rs/sadrzaj/2016/978-86-6065-397-2>
- [4] Киш, Н. (2021) Рекцијски речник придева. У С. Ристић, И. Лазић Коњик, Н. Ивановић (ур.) *Лексикографија и лексикологија у светлу актуелних проблема* (стр. 969–983). Институт за српски језик САНУ.
- [5] Mel'čuk, I. (1988). *Dependency Syntax: Theory and Practice*. Suny Series in Linguistics. State University Press of New York.
- [6] Пипер, П. и др. (2005). *Синтакса савременог српског језика. Проста реченица*. Институт за српски језик САНУ – Београдска књига – Матица српска.
- [7] Радовановић, М. и Бугарски, Н. (2007). Номинализације: два лика. У *Зборник Матице српске за славистику* 71–72 (стр. 199–209).
- [8] Штрбац, Г. (2010). Специјални граматички речници у лексикографији. У В. Ружић, Д. Маринковић (ур.) *Годишњак Филозофског факултета у Новом Саду XXXV/2* (стр. 369–379). Филозофски факултет.
- [9] Utvić, M. (2008). *Konačni automati u regularnoj imenskoj derivaciji* (nepublikovana magistarska teza). Univerzitet u Beogradu chromeextension: //efaidnbmnnnibpccajpcgclefindmkaj/<https://informatika.fil.bg.ac.rs/misko/bibliografija/2008/msMilosUtvic.pdf>
- [10] Utvić, M. (2014). *Izgradnja referentnog korpusa savremenog srpskog jezika* [Doktorska disertacija, Univerzitet u Beogradu]. [NaRDuS] <https://nardus.mpin.gov.rs/handle/123456789/4091>

Syntactic-semantic Annotation of Serbian Language Electronic Corpora

Nataša Kiš

Summary

This paper explores the potential of using electronic corpora as a primary source of data for various studies in the fields of syntax and semantics of contemporary Serbian. It has two main objectives. The first is to present previous research on the complementation of adjectives and adjectival nouns – specifically, their government potential at the syntagmatic level – based on relevant material drawn primarily from the *Corpus of Contemporary Serbian Language*, SrpKor2013 (www.korpus.matf.bg.ac.rs).

The second objective is to emphasize the need to expand the search capabilities of electronic corpora to support syntactic-semantic analysis at the level of the simple sentence. Such searches rely primarily on syntactic and semantic corpus annotation and would include, among other features, data on the valency properties of predicates and on both obligatory and optional arguments within specific contexts. A key component of this process is the classification of semantic word classes and their syntactic-semantic properties, which would enable targeted searches for different sentence models.

One of the paper's goals is to outline the types of linguistic analyses required to extract and integrate relevant data for the further development of syntactic and semantic corpus annotation. The study aims to demonstrate the complexity of these annotation processes, which necessitate solid theoretical foundations specific to each natural language. Even within the same language families, grammars may vary significantly at the syntactic and semantic levels – differences that directly impact text processing and parsing strategies.

Therefore, collaboration between linguists and computational experts is essential to ensure that linguistic insights inform more accurate and efficient annotation. At the same time, the application of syntactically and semantically annotated corpora in linguistic research and language education – whether in teaching Serbian as a first, second, or foreign language – could substantially enhance the effectiveness and scope of both research and pedagogical practices.

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

Značaj paralelnih korpusa za istraživanje frazemskih konstrukcija u nemačkom i srpskom jeziku

Naučni rad

DOI: 10.18485/judig.2025.1.ch21

Kristina Ilić¹,  0009-0002-7426-5530

Apstrakt

Paralelni korpusi su od velikog značaja za kontrastivna istraživanja frazemskih konstrukcija u različitim jezicima. Prema Dobrovoljskom (2011, str. 114), frazemske konstrukcije definisane su kao leksički delimično specifični parovi oblika i značenja, čija semantika proizlazi ne samo iz leksičkog značenja njihovih komponenti i načina njihove sintaktičke povezanosti, već i iz same konstrukcije kao celine. Frazemske konstrukcije su idiomatske kombinacije reči koje se sastoje od određenih leksičkih elemenata, tzv. sidra i određenog broja praznih mesta, tzv. slotova. Ove praznine popunjavaju se tzv. filerima, pri upotrebi konstrukcije (Dobrovol'skij 2011, str. 114). Konstrukciona gramatika je mlada grana lingvistike, a istraživanje frazemskih konstrukcija uz pomoć korpusa, naročito paralelnih korpusa u slučaju kontrastivnih istraživanja, sve više se razvija. U ovom trenutku može se reći da su paralelni korpusi nemačkog i srpskog jezika slabije zastupljeni, za razliku od paralelnih korpusa nemačkog i drugih većih svetskih jezika ili paralelnih korpusa drugih većih svetskih jezika i srpskog jezika. Cilj ovog rada je predstavljanje do sada dostupnih resursa u vidu paralelnih korpusa nemačkog i srpskog jezika i mogućih načina njihove upotrebe u kontrastivnom istraživanju frazemskih konstrukcija na konkretnim primerima, a potom i predlog optimalnog resursa za kontrastivno istraživanje frazemskih konstrukcija u nemačkom i srpskom jeziku.

¹ Filološki fakultet Univerziteta u Beogradu, kristina.ilic997@gmail.com

Ključne reči: Konstrukciona gramatika, korpusna lingvistika, frazemske konstrukcije, srpski jezik, nemački jezik

1. Uvod

Ovaj rad zasnovan je na sponi između konstrukcione gramatike i korpusne lingvistike i prikazuje praktičnu upotrebu paralelnih korpusa iz perspektive krajnjeg korisnika, koji sebi postavlja kao cilj istraživanje frazemskih konstrukcija.

Teorijsku osnovu rada sa jedne strane čine postulati konstrukcione gramatike i frazemske konstrukcije kao jedan od njenih predmeta istraživanja, dok sa druge strane rad predstavlja tipologiju i pruža opis različitih postojećih paralelnih korpusa, sa fokusom na paralelne korpusne srpskog i nemačkog jezika (*SrpNemKor*, *InterCorp* i *OpenSubtitles 2018 parallel – German*) i srpskog i italijanskog (*SerbItaCor3*).

Različiti nivoi složenosti frazemskih konstrukcija zahtevaju korišćenje upita različitog nivoa kompleksnosti za pretragu korpusa. U ovom radu će biti obrađena tri nivoa složenosti frazemskih konstrukcija i upita pomoću kojih se one pretražuju u korpusu.

Empirijsko istraživanje sastoji se od opisa različitih mogućnosti pretrage korpusa korišćenjem konkretnih primera frazemskih konstrukcija, u cilju predstavljanja postojećih i nedostajućih potrebnih funkcionalnosti korpusa, kako bi se dobili optimalni rezultati.

Cilj ovog rada je dati predlog optimalnog resursa za kontrastivno istraživanje frazemskih konstrukcija u nemačkom i srpskom jeziku, u pogledu sadržaja i funkcionalnosti.

2. Paralelni korpusi

Predmet istraživanja korpusne lingvistike predstavljaju korpusi različitog obima i sadržaja, koji su pretraživi uz pomoć korpusnih menadžera, različitih funkcionalnosti. Korpusni menadžeri predstavljaju kompjuterske programe koji omogućavaju izgradnju, uređivanje, obeležavanje tj. anotaciju i pretraživanje korpusa².

Paralelni korpusi predstavljaju jednu od vrsta korpusa. „Paralelni korpusi (poravnati korpusi) sastoje se od nekoliko semantički ekvivalentnih tekstova, koji su poravnati na određenom nivou — na nivou pasusa, rečenica, fraza ili reči.“ (Krstev & Vitas, 2011, str. 495). Svrha korpusa određuje: (1) izbor tekstova, (2) njihovu veličinu, (3) nivo poravnanja i (4) metod korišćen za poravnanje (Krstev & Vitas, 2011, str. 495–496). Autori ovog određenja

² <https://www.sketchengine.eu/glossary/corpus-manager/>, pristupljeno 07.11.2025.

svrhe korpusa, Krstev i Vitas, naglašavaju da poravnati korpus ne mora biti multilingvalan, već se može sastojati i iz dva ili više nezavisnih prevoda istog teksta na jednom jeziku (Krstev & Vitas, 2011, str. 495).

Blumel, citirajući McEnery & Xiao, prevodne korpuse (eng. *translation corpora*) deli na (1) paralelne korpuse, koji se sastoje od izvornih tekstova i njihovih prevoda na jedan ili više drugih jezika i oni mogu biti (a) jednosmerni (eng. *unidirectional*), (b) dvosmerni (eng. *bidirectional*) i (c) višesmerni (eng. *multidirectional*); zatim na (2) uporedne korpuse (eng. *comparable corpora*), koji se sastoje od sličnih tekstova i na (3) ekvivalentne korpuse, koji predstavljaju kombinaciju prva dva tipa (Blumel, 2014, str. 32, citirano prema McEnery & Xiao, 2008).

Uporedni korpusi se ne sastoje nužno od semantički ekvivalentnih tekstova, već od tekstova koji pripadaju istom domenu, istom vremenskom periodu itd.“ (Krstev & Vitas, 2011, str. 496).

U pogledu sadržaja, „Paralelni korpusi kao kolekcije tekstova na dva ili više jezika najčešće sadrže administrativne, pravne ili novinske tekstove, te zapravo predstavljaju korpuse podjezika”, pri čemu su „višejezični paralelni korpusi književnih tekstova najčešće mnogo manji u odnosu na korpuse pravnih ili novinskih tekstova jer njihova priprema zahteva više vremena i rada, ali je i njihov broj poslednjih godina u porastu” (Andonovski, 2019, str. 29–30).

Kako navode Doval i Sančez Nieto, paralelni korpusi se istražuju najpre kao proizvodi, ali se može izučavati i proces njihove izrade, koji podrazumeva „kompilaciju, anotaciju, poravnanje, indeksiranje i vršenje upita nad korpusom”³ (Doval & Sánchez Nieto, 2019, str. 6–17).

Korpusi, pored toga što mogu predstavljati sam predmet istraživanja, kao u slučaju korpusne lingvistike, služe i kao alati koji se koriste za sprovođenje različitih lingvističkih istraživanja. Andonovski (2019), citirajući Dobrića, navodi različite oblasti lingvistike u kojima korpusi nalaze široku primenu, kao što su morfologija, sintaksa, semantika, diskurs, pragmatika, leksikologija, leksikografija (Andonovski, 2019, str. 10–11, prema Dobrić 2009, str. 362), pri čemu naglašava koje sve mogućnosti korpusi nude.

Kako navodi Andonovski, u zavisnosti od nivoa anotacije, korpusi omogućavaju istraživačima da uoče različite primere upotrebe jezika, da pozivanjem seta reči jednog jezika i njihovih gramatičkih oblika prilikom istraživanja sagledaju frekventnost pojavljivanja određene reči ili fraze postavljene kroz upite za pretragu, u kojim se sve oblicima i varijantama postavljena reč ili fraza pojavljuje, kao i u kakvoj je semantičkoj korelaciji

³ Svi prevodi direktnih citata u radu su od strane autora.

sa drugim rečima i frazama i njihovim oblicima u datom korpusu (Andonovski, 2019, str. 11).

Navedene funkcionalnosti pretrage korpusa pokazaće se kao izuzetno značajne za korpusna istraživanja frazemskih konstrukcija.

Paralelni i uporedni korpusi, prema Doval i Sančez Nieto (Doval & Sánchez Nieto, 2019, str. 4), nalaze primenu u (1) mašinskom prevodenju, i to u vidu podataka za trening prevodilačkih modela, prevodnih memorija ili leksičku i terminološku ekstrakciju, u višejezičnoj obradi prirodnih jezika (NLP) i u (2) lingvističkim istraživanjima u oblasti kontrastivne lingvistike i translatologije, prevodilačke prakse, leksikografije i oblasti podučavanja stranih jezika i prevodenja.

Analiza dvojezičnog korpusa, prema Modercu (2015), može se fokusirati na traduktološke i stilističke aspekte pojedinih prevoda ili više prevoda istovremeno, pratiti rad pojedinih prevodilaca ili porediti prevode istog dela od strane različitih prevodilaca (Moderc, 2015, str. 301).

Andonovski, citirajući Simeona, ističe značaj paralelnih korpusa za višejezičnu leksikografiju, i to u tome što „omogućavaju korisnicima da sagledaju uparene ekvivalente određenih segmenata teksta kao i širi kontekst koji ne postoji u klasičnim dvojezičnim rečnicima” (Andonovski, 2019, str. 24, prema Simeon 2002: 212).

Tvorci višejezičnog paralelnog korpusa *InterCorp*, Čermák & Rosen (Čermák & Rosen, 2012, str. 1), naglašavaju značaj paralelnih korpusa „posebno za jezike sa manjim brojem izvornih govornika, podržavajući i uporedna istraživanja i proučavanja jednog jezika iz perspektive drugih jezika”.

3. Postojeći paralelni korpusi srpskog i drugih jezika

S obzirom na činjenicu da se radi o jeziku sa malim brojem govornika tj. srpskom, a izrada korpusa predstavlja složen zadatak, paralelni korpusi srpskog i drugih svetskih jezika nisu brojni, kao ni opšti korpusi srpskog jezika, ali u poslednjim godinama beleži se njihov porast. Paralelni korpusi značajni za ovo istraživanje su:

- (1) *SrpNemKor*, dostupan na platformi *Biblisha*⁴, paralelni korpus nemačkog i srpskog jezika koji se sastoji od po 7 romana na oba jezika i njihovih prevoda, nastao u okviru doktorske disertacije *Mreža otvorenih podataka i jezički resursi u procesu izgradnje srpsko-nemačkog literarnog korpusa* (Andonovski, 2019);

⁴ <http://biblisha.jerteh.rs/>, pristupljeno 08.08.2025.

- (2) *InterCorp*, dostupan preko korpusnog menadžera *KonText*⁵, pri čemu je fokus na podkorpusu *DE-SR* iz centralnog dela korpusa, pod nazivom *Core* (Rosen, 2016);
- (3) *SerbItaCor3*, dostupan na platformi *NoSke*⁶, paralelan korpus italijanskog i srpskog jezika, sačinjen od 243 tekstova, odnosno 267 uparenih tekstova (Moderc et al., 2023);
- (4) *OpenSubtitles 2018 parallel – German*, koji sadrži prevode filmova i deo je kolekcije *OpenSubtitles 2018 parallel*⁷, sačinjene od 60 korpusa na 58 jezika, pohranjen na korpusnom menadžeru *Sketch Engine* (Lison & Tiedemann, 2016).

U Tabeli 1 dat je uporedni prikaz ovih korpusa, prema jezicima, izvorima, nivou poravnanja, podržanosti *CQL-a*⁸ i funkcionalnostima korpusnog menadžera.

Tabela 1: Korpusi obuhvaćeni istraživanjem

korpus (platforma) i autor	jezici	izvori	poravnanje	CQL	najvažnije funkcionalnosti
<i>SrpNemKor</i> (Biblisha) Jelena Andonovski	srpski i nemački	7 srpskih romana 7 nemačkih romana (različitih)	nivo rečenice	ne	smer srpski → nemački
<i>SerbItaCor3</i> (NoSke) Saša Moderc	srpski i italijanski	267 prevoda ukupno (ima i više prevoda jednog književnog dela)	nivo rečenice	da	parallel concordance, concordance, word list, keywords, frequency text type analysis
<i>InterCorp</i> (KonText) Alexandr Rosen	nemački i srpski	core (književna dela, stručne knjige) collections (<i>Syndicate</i> , <i>Acquis</i> , <i>Europarl</i> , <i>Open Subtitles</i> , Biblija)	nivo rečenice prema češkom	da	parallel concordance, word list, keywords, frequency, collocations,
<i>OpenSubtitles 2018 parallel – German</i> (Sketch Engine) Jörg Tiedemann	engleski ↔ nemački ↔ srpski	titlovi za film sa www.OpenSubtitles.org	nivo rečenice prema engleskom	da	parallel concordance concordance word list keywords frequencyword sketch, thesaurus, n-grams,

⁵ <https://intercorp.korpus.cz/>, pristupljeno 08.08.2025.

⁶ <https://noske.jerteh.rs/>, pristupljeno 08.08.2025.

⁷ <https://www.sketchengine.eu/opensubtitles-parallel-corpora/>, pristupljeno 08.08.2025.

⁸ *Korpusni jezik za upite (CQL)* je poseban kod ili jezik za upite koji se koristi [...] za pretragu složenih gramatičkih ili leksičkih obrazaca ili za korišćenje kriterijuma pretrage koji se ne mogu podesiti pomoću standardnog korisničkog interfejsa.

<https://www.sketchengine.eu/documentation/corpus-querying/>, pristupljeno 08.08.2025.

				text type analysis

4. Frazemske konstrukcije

„Frazemske konstrukcije se mogu definisati kao konstrukcije koje kao celina imaju leksičko značenje, pri čemu su određene pozicije u njihovoj sintaksičkoj strukturi popunjene leksičkim elementima, dok druge predstavljaju slotove (eng. Slots), koji moraju biti popunjeni“ (Dobrovol'skij, 2011, str. 114).

Prema Dobrovol'skom, frazemske konstrukcije su jedan od podtipova frazema. Ovaj autor deli frazeme na (1) idiome, koji pokazuju „izrazite nepravilnosti u odnosu forme i značenja“, kao npr. *ins Graß beißen* (srp. *umreti; otegnuti papke*); (2) kolokacije, koje se sastoje od osnovne konstituente sa punim značenjem i vezane komponente sa oslabljenim ili metaforičkim značenjem, u koje spadaju i funkcionalni glagolski spojevi, kao npr. *eine Entscheidung treffen* (srp. *doneti odluku*); (3) situacione klišeje tj. rutinske formulacije, npr. *Guten Tag!* (srp. *Dobar dan!*); (4) izreke, npr. *Morgenstund hat Gold im Mund.* (srp. *Ko rano rani, dve sreće grabi.*); (5) gramatičke frazeme, npr. *geschweige denn* (srp. *a kamoli*) i (6) frazemske konstrukcije tj. frazeošablone, kao na primer *So ein Glück!*, sa apstraktnom shemom/šablonom (nem. *Pattern*) [*so EIN N*]⁹ (Dobrovol'skij, 2011, str. 111).

To su stabilne višečlane jedinice, koje se sastoje od popunjenih delova, tzv. sidra (nem. *Anker*) i slobodnih slotova (nem. *Slots*), koji se prilikom upotrebe popunjavaju leksemama ili kombinacijama leksema koje se nazivaju fileri (nem. *Filler* ili *Slotfüller*) (Pavlova, 2022, str. 594).

Prema Ilić i Šlund (2024), „do ekvivalenata FK [frazemskih konstrukcija] može se doći na nekoliko različitih načina, u zavisnosti od dostupnosti resursa: (1) pretragom u paralelnim korpusima nemačkog i ciljnog jezika; (2) u prevodima književnih dela sa nemačkog govornog područja na ciljni jezik, ručno ili uz pomoć paralelnih korpusa književnih dela; ili (3) samostalnim prevodom“ (Ilić & Šlund, 2024, str. 291). U ovom radu ćemo se fokusirati na prva dva slučaja, jer je fokus rada na upotrebi korpusa.

Prema složenosti upita za pretragu u jednojezičnom korpusu, u svojoj prezentaciji Krotova (2024) razlikuje tri nivoa složenosti frazemskih konstrukcija. Na *Niveau 1* nalaze se frazemske konstrukcije sa malo ili bez ograničenja kod filera, malo ili bez varijacije kod sidra, te za istraživanje nije nužno koristiti anotirane korpusne, već su dovoljni pretraga na Guglu,

⁹ N označava da na tom mestu stoji imenica.

onlajn rečnici i *Google Books Ngram-Viewer*¹⁰. Primeri takvih frazemskih konstrukcija su *es sei denn X* (srp. *sem/osim ako ne*) i *ob X oder nicht* (srp. *bio/bila/bilo X ili ne*). Za frazemske konstrukcije na Nivou 2 preporučljivo je koristiti korpusne alate sa anotiranim korpusima koji sadrže anotaciju vrsta reči, tzv. *POS-Tagging*, i mogućnost navođenja željenog razmaka između elemenata u okviru upita¹¹, kako bi upit bio precizan. Frazemska konstrukcija kod koje je takav postupak potreban je, na primer, *du und dein X/ du mit deinem X* (srp. *ti i tvoj/tvoja/tvoje/ tvoji X*). Za istraživanje frazemskih konstrukcija na Nivou 3 potrebno je da korpusni alati imaju mogućnost pretrage ponavljanja jednog elementa i da omogućavaju generisanje frekvencijskih lista za filere, kao što je slučaj kod korpusnog menadžera *Sketch Engine*. Alternativa bi bili regularni izrazi ili programski jezik *Python* za generisanje frekvencijskih lista. Ovom nivou, na primer, pripada frazemska konstrukcija *Kind ist Kind* (srp. *dete je dete*) sa apstraktnom shemom *X' ist/sind [immer] X' (X' je/su X')* (Krotova, 2024). Ove klasifikacije ćemo se pridržavati i u ovom radu.

5. Frazemske konstrukcije i pretraga u paralelnom korpusu: rezultati istraživanja

Rezultati istraživanja pokazuju da se frazemske konstrukcije u nemačkom jeziku mogu efikasno identifikovati i analizirati u paralelnim korpusima, ali da način njihove pretrage zavisi od formalne strukture same konstrukcije. Stoga je važno da korpusni menadžeri poseduju određene tehničke funkcionalnosti. Konstrukcije su klasifikovane na osnovu stepena složenosti, uzimajući u obzir tehničke funkcionalnosti i složenost upita potrebnih za pretragu u korpusu.

Nivo 1. Za analizu konstrukcija koje imaju jednostavno sidro i jedan slot sa velikim brojem mogućih filera, poput *es sei denn X* (srp. *osim ako ne X*) ili *durch und durch X* (srp. *skroz na skroz X*), odnosno *X durch und durch* (srp. *X skroz na skroz*), moguće je koristiti jednostavne metode pretrage u paralelnim korpusima. Za njihovu pretragu nije potrebna POS anotacija i dovoljno je u polje za pretragu uneti sidro *durch und durch*. U korpusu paralelnih tekstova se vidi da li je i u drugom jeziku upotrebljena frazemska konstrukcija.

S obzirom na to da je za *SrpNemKor* na platformi *Biblisha* moguća samo pretraga u smeru srpski → nemački, moguće je doći do prevodnih ekvivalenata za konstrukciju *durch und durch X* (srp. *skroz na skroz X*) tako

¹⁰ <https://books.google.com/ngrams/>, pristupljeno 15.07.2025.

¹¹ Na primer, navođenje {0,2} posle elementa frazemske konstrukcije u okviru upita znači da se naredni element nalazi odmah pored, nakon jednog tokena ili nakon dva tokena.

što se postavi upit koji sadrži njen mogući prevod. Pogodak je pronađen uz pomoć jednog od sinonima izraza *skroz na skroz*, odnosno uz pomoć reči *potpuno*, kao što se vidi na Slici 1, što znači da se prevodilac nije opredelio za frazemsku konstrukciju u srpskom jeziku, već je posegnuo za prilogom sa sinonimnim značenjem.

Das Parfum / Patrick Suiskin = Parfum / Patrick Ziskind, ID: 11.1.007 metadata	n2174 Sie mussten durch und durch getränkt sein von seinem Geruch.	n2174 Mora da su potpuno natopljene njegovim mirisom.

Slika 1: Pretraga u korpusu SrpNemKor na osnovu reči „potpuno“

Unošenjem sidra u polje za pretragu može se vršiti pretraga i u korpusima koji se nalaze na platformama koje se baziraju na korpusnim menadžerima *NoSketchEngine* i *SketchEngine*, odnosno u korpusu *SerbItaCor3*, kao što se vidi na Slici 2 i u korpusu *OpenSubtitles 2018 parallel – German*, prikazanom na Slici 3¹². Moguće je analizirati frekvenciju filera ili slotova koji se nalaze u pretraživanoj konstrukciji. Ukoliko je potrebno analizirati frekvenciju u ciljnom jeziku, to je moguće samo unošenjem ekvivalentne konstrukcije i u ciljnom jeziku u pretragu¹³. U tom slučaju dobijamo samo rezultate u kojima je konstrukcija prevedena odgovarajućom konstrukcijom. Nije moguće dobiti frekvenciju različitih ekvivalenata konstrukcije u ciljnom jeziku, što bi predstavljalo značajan podatak.



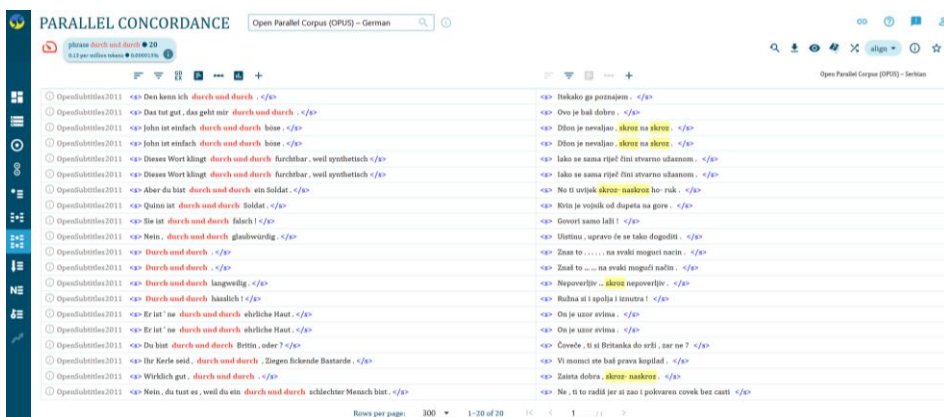
Slika 2: Pretraga konstrukcije „skroz naskroz X/X skroz naskroz“ u korpusu SerbItaCor3.

Na isti način vrši se i pretraga u korpusu *InterCorp* preko platforme *KonText*. U rezultatu, obeležavanjem bilo koje reči u izvornom tekstu, ne samo reči obuhvaćenih pretragom, automatski se istom bojom označi ekvivalent te reči u ciljnom tekstu, ukoliko ga ima, kao što se vidi na Slici

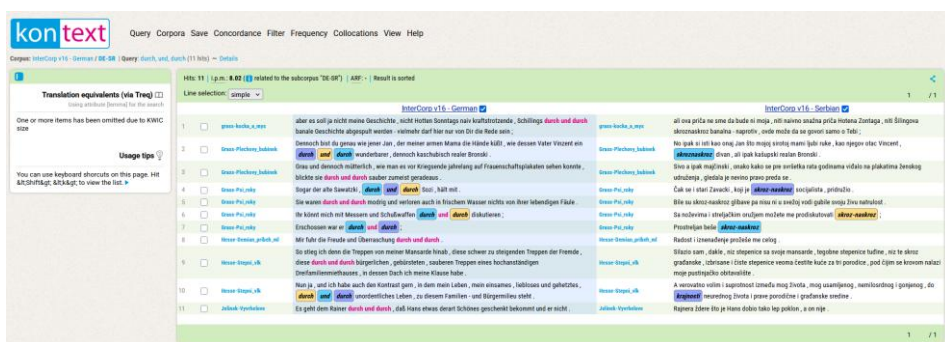
¹² Korpus *OpenSubtitles 2018 parallel*, kako se vidi na Slici 3, ima sledeći nedostatak: paralelni tekstovi ne predstavljaju nužno izvorni tekst i njegov prevod, već je kod oba teksta često u pitanju prevod sa engleskog. To otežava lingvistička i translatološka istraživanja.

¹³ <https://www.sketchengine.eu/guide/parallel-concordance-searching-translations/#toggle-id-2>, pristupljeno 07.08.2025.

4. Direktnog ekvivalenta pojedinačne reči nema ukoliko je prevodilac izabrao neko drugo rešenje. Dolazi i do grešaka u obeležavanju. Na platformi *Sketch Engine* nije moguće kliknuti na pojedinačne reči i dobiti obeležene ekvivalente.



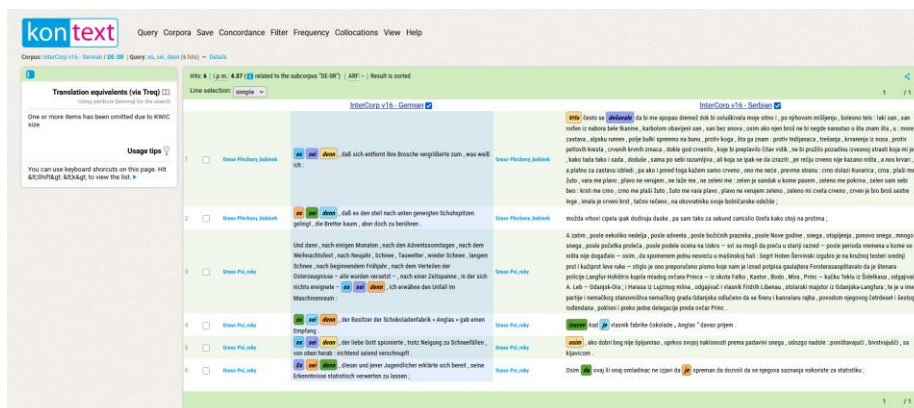
Slika 3: Pretraga konstrukcije „durch und durch X/X durch und durch“ u korpusu OpenSubtitles



Slika 4: Pretraga konstrukcije „durch und durch X/X durch und durch“ u korpusu InterCorp

Primeri pogrešnog ili izostalog obeležavanja ekvivalenata u korpusu *InterCorp* prikazani su na Slici 5, na primeru frazemske konstrukcije *es sei denn X* (srp. *osim ako ne X*).

Nešto složenija za pretragu je konstrukcija *ob X oder nicht*, s obzirom na to da se filer nalazi između dva dela sidra. Stoga bi trebalo upotrebiti *CQL* upit `[word="ob"] [word="oder"] [word="nicht"]` kako bi se dalo uputstvo da je mesto slotu između dva dela sidra, odnosno posle reči *ob*.

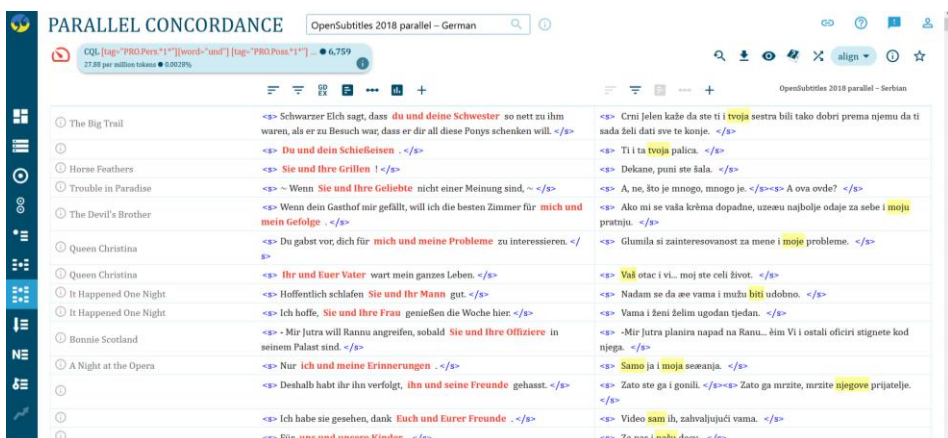


Slika 5: Primeri pogrešnog ili izostalog obeležavanja ekvivalenata u korpusu InterCorp

Nivo 2. Kod konstrukcija koje uključuju lične i posesivne zamenice u sidru, pretraga postaje znatno složenija. Takve su konstrukcije *du und dein X* i njene varijante *du [schon wieder/immer] mit deinem X* i *du mit deinem X [schon wieder?/!]*, čije apstraktne sheme glase $Pron_{pers} \text{ nom}/N'_{nom}$ und dein(e)/sein(e)/ihr(e)/euer(eure)... N'_{nom} , odnosno $Pron_{pers} \text{ nom}/N'_{nom}$ [schon wieder/immer] mit deinem/seinem/deiner/seiner/seinen/ihr(e)em... N'_{nom} [schon wieder?/!], čiji bi srpski ekvivalent bio *Ti i tvoja/tvoja/tvoje N_{nom}*. Zamenice i glagol mogu stajati u bilo kojem licu. Kod ovakvih konstrukcija, neophodno je koristiti anotirane korpuse sa POS oznakama.

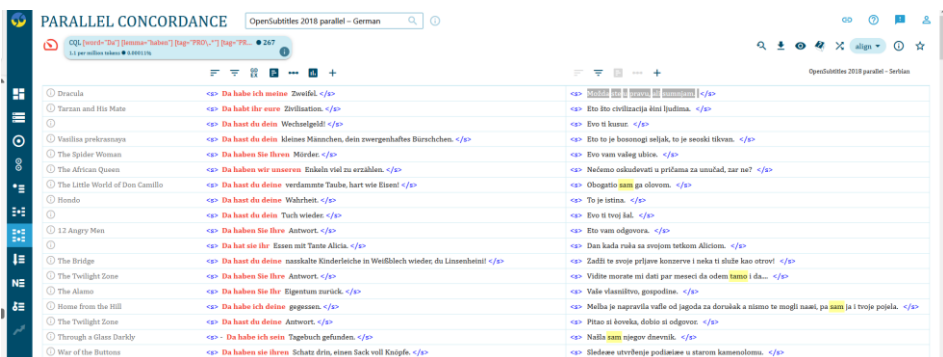
Kada se istražuje konstrukcija *du und dein X* u korpusnim menadžerima koji podržavaju CQL, ukoliko se postavi opšti upit [*tag="PRO\..*" | tag="N.Name.Nom.*" | tag="N.Reg.Nom.*"*] [*word="und"*] [*tag="PRO.Poss.*"*] [*tag="N.*"*], koji bi na prvi pogled obuhvatio sva moguća lica i dao najbolje rezultate, dobije se veliki broj rezultata koji nisu frazemska konstrukcija, već obuhvataju različita lica lične i posesivne zamenice ili nemaju značenje konstrukcije, već predstavljaju složeni subjekat, kao npr. *Nora und ihr Mann kommen heute Abend*. (srp. *Nora i njen suprug dolaze večeras*).

Kako bi se suzila pretraga, potrebno je u okviru CQL upita precizirati da je kod lične zamenice i kod prisvojne zamenice u pitanju isto lice: [*tag="PRO.Pers.*1*"*][*word="und"*] [*tag="PRO.Poss.*1*"*] [*tag="N.*"*], pri čemu se dobiju rezultati prikazani na Slici 6. Ovom pretragom su obuhvaćena sva lica, ali su zamenice u pojedinačnim primerima u istom licu. Zatim je potrebno ručno izdvojiti primere u kojima se javlja tražena konstrukcija, kao što su *Du und dein Schießseisen* i ekvivalent *Ti i tvoja palica* sa Slike 6. Ova pretraga ne daje rezultate u korpusu *InterCorp*.



Slika 6: Rezultati pretrage [tag="PRO.Pers.*1*"] [word="und"]
[tag="PRO.Poss.*1*"] [tag="N.*"]

Takva je i konstrukcija Da hast du [dein/deinen/deine] N_{Akk} [!] (srp. Evo ti tvoj/tvoja N_{nom}), sa apstraktnom shemom Da habe/hast/hat/habt/haben Pron_{pers} nom [Pron poss akk] N_{Akk} [!]. Ona se u korpusu može pretraživati uz pomoć CQL upita [word="Da"] [lemma="haben"] [tag="PRO\..*"] [tag="PRO.Poss.*"], pri čemu veliki deo rezultata sadrži tu konstrukciju, kao što se vidi na Slici 7.



Slika 7: Rezultati pretrage [word="Da"] [lemma="haben"]
[tag="PRO\..*"] [tag="PRO.Poss.*"]

Zbog razlike u tagovima u različitim korpusima, potrebno je prilagoditi upit za pretragu u korpusu *InterCorp*, koji onda glasi [word="Da"] [lemma="haben"] [tag="P.*"] [tag="P.*"], a rezultati su prikazani na Slici 8. Među ovim rezultatima, frazemske konstrukcije su manje zastupljene nego u slučaju korpusa *OpenSubtitles*, s obzirom na to

da je korpus ograničen na 13 romana koji su prevedeni sa nemačkog na srpski. Osim toga, ova konstrukcija je karakteristična za govorni jezik, pa je zastupljenija u filmskim titlovima.

The screenshot shows the Kontext web interface. At the top, there's a search bar with the query 'da haben wir es also?'. Below the search bar, there's a table of results. The table has columns for 'Line selection', 'InterCorp v16 - German', and 'InterCorp v16 - Serbian'. The results are numbered 1 through 17. Each result has a checkbox and a text snippet. The snippets are in German and Serbian, showing various contexts where the phrase 'da haben wir es also?' is used.

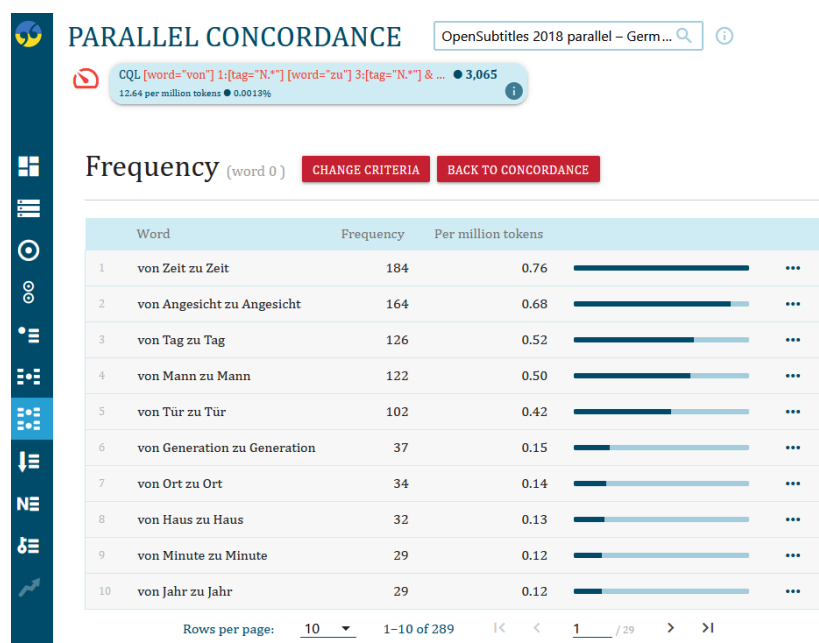
Line selection	InterCorp v16 - German	InterCorp v16 - Serbian
☐ 1	Da habe ich sie, die Trübsal.	I see you, he is, indeed.
☐ 2	Da hätte ich allen Mut zusammennehmen müssen, um als schlechter, verlässlicher Sohn meinen braven Kinde antworten zu können.	Tu bih morao da prikažem ovu svoju hladnost da kao rđav, izdajnički sin odgovorim ovom dobrom detetu.
☐ 3	Da hätte ich was nicht alles, womöglich einen Präzisten herbeiführen können.	Mogao bih diskovom da diskovom ne znam koga ovo, možda i jednjaka.
☐ 4	Da hätte ich meinen Auftritt in Heavilly Möbelpackerei duldet, indem ich mich selbst, in unseren Keller Platz zu schaffen für Ölgemälde und Ölgemälde, für Kunst und Musik.	Tada sam ja nastupio u kaveznoj (povremeno bledo nasača nemeštaja, jer sam se ponudio da u našem podrumu napravim mesto za bleđan grafi i bleđani grafi, bleđi i meštari).
☐ 5	Da hast du die aber einen anderen Vogel eingefangen.	Ti si tu udele odvojio (plovu i psi).
☐ 6	Da hast du die einen klischen Weg ausgepickt.	Smislilo si prepoznati put.
☐ 7	Da hätte er es seinen Herzen gefüllt.	Tada je svojim svojim srcem osvećao: - Prired tobom se našao put za koji si govorio, bogovi te otkriva.
☐ 8	Da hat er sich vor Sophie eigens im Anzug zeigen wollen, damit er einen Vorsprung auf den Rainer und Alexen Janssen hat, und jetzt wird er so ausgegottet!	Eto, on je sa sobom zbog Soti hteo da se pokazuje u odeću, da bi dobio malo prednosti u odnosu na Rainera i njegovu oticu, i sad brzo tako iznervira.
☐ 9	Da hat er die Gedanken, das Seine, und die persönliche Güter macht es komplett zunichte.	Upravo ga je to oduzelo, gumpu, a personalna svojstva je one uništio.
☐ 10	Da hast du mich für einen schönen Proben empfangen.	Za koga si me mislio prepoznati? *
☐ 11	Da habe ich ihm ein Glas durch die Luke gereicht.	A onda sam mu kroz otvor proterao čašu.
☐ 12	Da haben wir es also!	One su mi odmah pale u oči i sad sam te video prvi put, bilo je to pre nekoliko dana odprilike, onda, u kuzetiku? *
☐ 13	Da hast du's, Gottfried? "sagte Kister, das ist die Romanik von heute."	- Eto ti, Gottfried! - reče Kister. - To je današnja romanika!
☐ 14	Da haben wir es!	- Alas, to je!
☐ 15	Da hast du was zum Leben und hast deine Ruhe dabei.	Onda (od malih kuga da videli, a pri tom i svoj me).
☐ 16	Da haben wir es.	- Eto, videli.
☐ 17	Da hat er mich umgibt, daß ich ihn heute mitnehmen soll.	Preklopim mi je da ga poredem danas.

Slika 8: Rezultati pretrage [word="Da"] [lemma="haben"] [tag="P.*"] [tag="P.*"]

Nivo 3. Treći nivo složenosti obuhvata konstrukcije koje sadrže rekurentne elemente, odnosno kod kojih se ista reč, najčešće imenica, javlja kao filer u oba slotu. Tipični primeri uključuju konstrukcije poput *von Tag zu Tag* (srp. *iz dana u dan*) i *Kind ist Kind* (srp. *dete je dete*), odnosno apstraktne sheme *von X' zu X'* i *X' ist X'*. Za pretragu ovih konstrukcija u korpusnim menadžerima *Sketch Engine* i *NoSketch Engine* može se koristiti upit [word="von"] 1:[tag="N.*"] [word="zu"] 3:[tag="N.*"] & 1.lemma=3.lemma, pri čemu se dobiju rezultati sa Slike 9, odnosno 1:[tag="N.*"] [lemma="sein"] 2:[tag="N.*"] & 1.lemma=2.lemma, sa rezultatima sa Slike 10, pri čemu je izuzetno važno u upitu naznačiti koje reči su iste.

[illegible]

Frekvencijske liste u alatima poput *Sketch Engine* dodatno pomažu u identifikaciji najčešćih leksičkih jedinica koje se javljaju kao fileri. Metodološki, one zahtevaju kombinaciju korpusnih alata sa kvantitativnim analizama. Takva lista prikazana je na Slici 11.



Slika 11: Frekvencijska lista filera za konstrukciju von X' zu X'

U svim navedenim slučajevima, jasno je da izbor metode pretrage korpusa direktno utiče na uspešnost analize. Jednostavne konstrukcije mogu se pratiti čak i bez naprednih alata, dok fleksibilnije i kompleksnije strukture zahtevaju preciznu anotaciju i naprednu pretragu.

6. Zaključak

Rezultat ovog istraživanja predstavlja obrazloženi predlog funkcionalnosti koje treba da ima idealni korpusni alat za istraživanje frazemskih konstrukcija.

Na osnovu prikazanih karakteristika postojećih korpusa i opisa funkcionalnosti koje su potrebne za pretragu i istraživanje frazemskih konstrukcija, zaključuje se da bi idealan paralelni korpus nemačkog i srpskog jezika sadržao (1) romane na oba jezika paralelizovane na nivou rečenice, (2) POS anotaciju, (3) više različitih prevoda jednog književnog dela, (4) paralelizovane titlove, pri čemu je izvorni jezik nemački, a ciljani jezik srpski i obrnuto i (5) dodatno, ukoliko je moguće: ekvivalentni korpus koji bi sadržao i uporedni korpus sa novinskim tekstovima o događajima na oba govorna područja.

Idealan korpusni menadžer bi omogućavao prvo mogućnost pretrage u oba smera, drugo mogućnost pretrage uz pomoć CQL-a, treće

mogućnost selekcije bilo koje reči u konkordansi i automatsko obeležavanje njenog ekvivalenta u ciljnom tekstu i četvrto mogućnost generisanja frekvencijskih lista reči iz upita i rezultata.

Pretrage u paralelnim korpusima dragocene su za kontrastivne analize frazemskih konstrukcija i kontrastivna istraživanja uopšte. To je naročito značajno za jezike u čijim istraživanjima konstrukciona gramatika nije još razvijena, pa se ovakvim kontrastivnim istraživanjima doprinosi i razvoju nauke o ciljnom jeziku. Stoga je važno razvijati i unapređivati paralelne korpus srpskog i nemačkog, kao i srpskog i drugih jezika.

Zahvalnica

Ovo istraživanje sprovedeno je u okviru COST akcije PhraConRep Višejezični repozitorijum frazemskih konstrukcija na jezicima centralne i istočne Evrope („A Multilingual Repository of Phraseme Constructions in Central and Eastern European Languages”), koja ima za cilj da kontrastivno istraži frazemske konstrukcije nemačkog i srpskog jezika, pri čemu upotreba različitih alata i paralelnih korpusa, predstavljenih u ovom istraživanju, može odigrati važnu ulogu.

Literatura

- [1] Andonovski, J. S. (2019). *Mreža otvorenih podataka i jezički resursi u procesu izgradnje srpsko-nemačkog literarnog korpusa*. [Doktorska disertacija]. Beograd: Univerzitet u Beogradu, Filološki fakultet.
- [2] Blumel, B. (2014). *Learning in parallel: Using parallel corpora to enhance written language acquisition at the beginning level*. Dimension, 31–48.
- [3] Dobrovol'skij, D. (2011). Phraseologie und Konstruktionsgrammatik. In A. Lasch & A. Ziem (Hrsg.), *Konstruktionsgrammatik III. Aktuelle Fragen und Lösungsansätze* (S. 110–130). Stauffenburg.
- [4] Čermák, F. & Rosen, A. (2012). The Case of InterCorp, a multilingual parallel corpus. *International Journal of Corpus Linguistics*, 411–427.
- [5] Doval, I. & Sánchez Nieto, M. (2019). Parallel corpora in focus. An account of current achievements and challenges. U I. Doval, & M. Sánchez Nieto (ur.), *Parallel Corpora for Contrastive and Translation Studies. New resources and applications*, (str. 1–18). Amsterdam / Philadelphia: John Benjamins Publishing Company.
- [6] Krotova, E. (2024). Projektrelevante Webkorpora & Analysesoftware [Prezentacija]. *Training "Datenbearbeitung und Übersetzung im Repositorium auf GitHub" Osijek, Croatia. 4.–6. September 2024*. Preuzeto sa https://github.com/PhKW/PhKWB/blob/main/Workshops/Workshop_20240905.pdf

- [7] Krstev, C. & Vitas, D. (2011). An aligned English-Serbian corpus. U N. Tomović & J. Vujić (ur.), *ELLSIIR Proceedings (English Language and Literature Studies: Image, Identity, Reality)*, Volume I, 495–508. Belgrade: Faculty of Philology, University of Belgrade.
- [8] Pavlova, A. (2022). Mehrsprachige Datenbank der Phrasem-Konstruktionen. U A. Klosa-Kückelhaus, S. Engelberg, C. Möhrs, & P. Storjohann (ur.), *Dictionaries and society: Proceedings of the XX EURALEX International Congress* (str. 594–604). Mannheim: IDS.
- [9] Stanković, R., Krstev, C., Lazić, B. & Vorkapić, D. (2015). A bilingual digital library for academic and entrepreneurial knowledge management. *10th International Forum on Knowledge Asset Dynamics — IFKAD 2015*. Bari, Italy.
- [10] Илић К. & Шлунд, К., (2024). Увиди у изразу вишејезичног репозиторијума фраземских конструкција. У С. Марјановић (ур.) *Модерни речници у функцији просечнога корисника: стари проблеми, савремени правци и нови изазови*, Лексикографски сусрети, Вол. 1 (str. 285–296). Београд: Филолошки факултет Универзитета у Београду.

Korpusi

- [1] Andonovski, J. S. (2019). *Mreža otvorenih podataka i jezički resursi u procesu izgradnje srpsko-nemačkog literarnog korpusa*. [Doktorska disertacija]. Beograd: Univerzitet u Beogradu, Filološki fakultet.
- [2] Lison, P. & Tiedemann, J. (2016). OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. U Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., & Piperidis, S. *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC-2016)*(923–929). Portorož: European Language Resources Association (ELRA).
- [3] Moderc, S., Stanković, R., Tomašević, A. & Škorić, M. (2023). An Italian-Serbian sentence aligned parallel literary corpus. *Review of the National Center for Digitization*, 43, 78–91. <https://www.ncd.matf.bg.ac.rs/issues/43/09.pdf>
- [4] Rosen, A. (2016). InterCorp – A look behind the façade of a parallel corpus. U E. Gruszczyńska (ur.), *Polskojęzyczne korpusy równoległe: Polish-language parallel corpora* (str. 21–40). I Warszawa: Instytut Lingwistyki Stosowanej.

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

Kristina Ilić

Summary

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

phraseme constructions of German and of the Serbian language contrastively, whereby the use of different tools and parallel corpora, presented in this research, can play a crucial role in reaching that goal.

Key words: Construction grammar, corpus linguistics, phraseme constructions, Serbian language, German language

Нове унапређене верзије програма „Ка минималним паровима”

Научни рад

DOI: 10.18485/judig.2025.1.ch22

Данило Алексић¹,  0000-0002-1478-7487

Апстракт

Рад је посвећен новим верзијама једног кратког програма на Python-у за аутоматску ексцерпцију сегменталних минималних парова и парова сродних са сегменталним минималним паровима. (Сегменталним минималним паровима сматрају се они парови речи чији се чланови фонолошки разликују само по једном пару [сегменталних] фонема [односно по једном пару {сегменталних} фонемских низова или једном пару сачињеном од једне {сегменталне} фонеме и једног {сегменталног} фонемског низа] само на једној позицији [нпр. *е* и *и* у *лѐк* ~ *лѝк* {и *овн* и *ск* у *свѣтѡвнѧ* ~ *свѣтскѧ*}] или на више позиција [нпр. *л* и *р* у *мѡлѧр* ~ *мѡрѧл* {и *аз* и *ин* у *кѧзѧз* ~ *кѧнѧн*}). Паровима сродним са сегменталним минималним паровима сматрају се они парови речи чији се чланови фонолошки разликују само по прозодији и по једном пару [сегменталних] фонема [односно само по прозодији и по једном пару {сегменталних} фонемских низова или једном пару сачињеном од једне {сегменталне} фонеме и једног {сегменталног} фонемског низа] само на једној позицији [нпр. *ћ* и *ч* у *прѣгледаће* ~ *прегледа́че* {и *ац* и *ер* у *дестила́ција* ~ *дестилѣ́рија*}] или на више позиција [нпр. *е* и *и* у *бѣлѣ* ~ *бѣли* {и *ин* и *ос* у *лѧнѧн* ~ *лѡсос*}; уп. Алексић & Шандрих, 2021, pp. 569–570.) У најбржој досадашњој верзији овог програма (в. Алексић & Мркела, 2022) сегментални минимални парови и парови сродни са сегменталним минималним паровима образовани су помоћу речника, стандардног мапирајућег типа („standard mapping type”; Python

¹ Катедра за српски језик са јужнословенским језицима, Универзитет у Београду – Филолошки факултет, danilo.aleksic@fil.bg.ac.rs


```
"izACiJ|izACij|izAcIJ|izAcIj|izAciJ|izAcij|"
"izaCIJ|izaCIj|izaCiJ|izaCij|izacIJ|izacIj|"
"izaciJ|izacij)[A-Za-zĆ-ž-]*)(?![A-Za-zĆ-ž])"
```

према поднисици "izacij", и драстично успорава ионако трому прву верзију.

Брже су верзије 2 и 2.1 (Алексић & Мркела, 2022). (Верзију 2 је написао Данило Алексић, а упаривање токена у верзији 2.1 решење је Лазара Мркеле.) Оне кроз улазни корпус пролазе само једном, вадећи токене који садрже прву задату поднисуку, другу задату поднисуку или обе задате подниске. Задате подниске се у токенима записаним малим словима замењују специјалним карактером, "∇". На пример, ако су задате подниске "ovsk" и "sk", нисци "Harlemovskom", пошто се задате подниске преклапају, одговарају две поредбене ниске – "harlemov∇om" и "harlem∇om" – а нисци "harlemskom" одговара једна поредбена ниска: "harlem∇om". Главни део верзије 2 јесте Декартов производ торки које садрже токен са првом задатом подниском и торки које садрже токен са другом задатом подниском, нпр. Декартов производ торки ("ČAK", "čak", "∇ak") и ("vičan", "vičan", "vi∇an") на једној и торки ("Đak", "đak", "∇ak") и ("viđan", "viđan", "vi∇an") на другој страни, дакле нпр.:

```
{
  (
    ("ČAK", "čak", "∇ak"),
    ("Đak", "đak", "∇ak")
  ),
  (
    ("ČAK", "čak", "∇ak"),
    ("viđan", "viđan", "vi∇an")
  ),
  (
    ("vičan", "vičan", "vi∇an"),
    ("Đak", "đak", "∇ak")
  ),
  (
    ("vičan", "vičan", "vi∇an"),
    ("viđan", "viđan", "vi∇an")
  )
}
```

ако су задате поднискe "č" и "đ" и ако нема других токена са задатим поднискaма. У верзији 2 се из израчунатог Декартовог производа узимају међусобно различити токени које прате исте поредбене ниске. Главни део верзије 2.1 јесте речник (в. одељак „Новине у КаМП-у 2.2 и у КаМП-у 2.3”) у чијим су вредностима токени са другом задатом подниском, а чији су кључеви поредбене ниске. У верзији 2.1 међу тим кључевима траже се поредбене ниске токена са првом задатом подниском. И верзија 2 и верзија 2.1 улазни корпус токенизују помоћу речника. У верзије 2 и 2.1 уведен је и мод у којем се не занемарују разлике између великих и малих слова код ексцерпираних токена (прва верзија је дотичне разлике игнорисала).

У одељцима који следе представља се један тип могућности, оличен у КаМП-у 2.2 и КаМП-у 2.3, да се до сада најбржи КаМП додатно убрза.

Сви КаМП-ови су писани на програмском језику Python и за потребе овог истраживања покретани на његовим верзијама 3.8.2 и 3.13.5.

2. Зашто нове верзије?

У једној књизи о Python-у срећу се следеће смернице:

Now that you’ve made your code work, and possibly made it better than the initial version, it may be time to make it faster. Then, again, it may not. One very important rule (along with such principles as KISS = Keep It Small and Simple, or YAGNI = You Ain’t Gonna Need It) that you should heed when tempted to fiddle with your code to speed it up:

Premature optimization is the root of all evil.

— Donald Knuth, paraphrasing C. A. R. Hoare

Another way of stating this, in the words of Ken Thompson, co-inventor of UNIX, is “When in doubt, use brute force.” In other words, don’t worry about fancy algorithms or clever optimization tricks if you don’t really, really need them. If the program is fast enough, chances are that the value of clean, simple, understandable code is much higher than that of a slightly faster program. After all, in a few months, faster hardware will probably be available anyway.

But if you do need to optimize your program, because it simply isn't fast enough for your requirements, you absolutely should profile it before doing anything else. That is because it's really hard to guess where the bottlenecks are, unless your program is really simple. And if you don't know what's slowing down your program, chances are you'll be optimizing the wrong thing. (Hetland, 2008, pp. 362–363)

У другој се каже:

This chapter covers the subjects in the natural order in which they occur in development: testing first and foremost, debugging next, and optimizing last. Most programmers' enthusiasm focuses on optimization: testing and debugging are often (wrongly, in our opinion) perceived as being chores, while optimization is seen as being fun. ... the Pythonic approach to optimization—close to Jackson's classic "Rules of Optimization: Rule 1: Don't do it. Rule 2 (for experts only): Don't do it *yet*." (Martelli et al., 2017, p. 453)

У трећој: „Optimization is the altar where maintainability is sacrificed” (Ramalho, 2015, p. 91). Из перспективе датих цитата, КаМП 2.2 и КаМП 2.3 резултат су неодговорног располагања ресурсима. Наиме, (1) већ је постојала задовољавајуће ефикасна верзија, КаМП 2.1, (2) Лазар Мркела је као проблем у вези са брзином КаМП-а 2 који треба најпре решити препознао споро учитавање корпуса³, (3) исти аутор је као први корак у убрзавању КаМП-а 2 предложио додавање опције за креирање речника од корпуса који би се сачувао на диску, а као следећи могући корак поменуо паралелизацију претраге (Алексић & Мркела, 2022, p. 19; како се види из одељка „Новине у КаМП-у 2.2 и у КаМП-у 2.3”, КаМП 2.2 и КаМП 2.3 не остварују те циљеве) и (4) начин на који је у КаМП-у 2.2 и КаМП-у 2.3 постигнуто убрзање није осмишљен на основу профилисања.

У одговор би се могао цитирати Рејмонд Хетинцер: „There must be a better way!” (нпр. PyData, 2023, 12:03; SF Python, 2016, 9:21). Аутори који помажу програмерима да буду бољи у свом послу не сносе ризике преране и непотребне оптимизације који сносе програмери при развијању кода у некој софтверској компанији. Ако је контекст предавање о Python-у, сасвим је очекивано да се промовишу ефикаснији и лепши приступи. На пример, Хетинцер је током једног предавања о Python-у (Next Day Video, 2013, 12:50) питао публику како треба конкатенирати ниске и како ниске не треба конкатенирати иако и тај други, не препоручени приступ постиже исти циљ. Из

³КаМП-ови корпус учитавају у деловима чију величину и сепараторе бира корисник (Алексић & Мркела, 2022, pp. 15, 20–21; Алексић & Шандрих, 2021, pp. 572, 581).

перспективе учења, или из академске перспективе, израда бржих верзија КаМП-а и писање чланка о њима ипак нису губљење времена.

3. Новине у КаМП-у 2.2 и у КаМП-у 2.3

Кôд КаМП-а 2.2 и кôд КаМП-а 2.3 биће објављени на сајту аутора (Алексић, n.d.).

Главна разлика између КаМП-а 2.2 и КаМП-а 2.3 с једне и КаМП-а 2.1 с друге стране јесте употреба класа уместо неких речника. Уместо кључева и вредности, које имају речници, употребљавају се називи класовних атрибута, односно вредности класовних атрибута:

```
rečnik[reč] = reč.casefold()
"""Из токенизације КаМП-а 2.1."""

setattr(self, reč, reč.casefold())
"""Из токенизације КаМП-а 2.3."""
```

Према Бизлију (2006, р. 143), `setattr(object, name, value)` исто је што и `object.name = value`. Откуда онда избор уграђене функције `setattr()`, односно уграђених функција `getattr()` и `hasattr()` (уп.:

```
class Rečnik2:
    def __init__(self):
        for torka_ in torka_2[0]:
            if not hasattr(
                self, torka_[0]):
                setattr(
                    self, torka_[0], {})
            getattr(self, torka_[0])[
                torka_[1]] = torka_[2]
```

), уместо нотације тачком (нпр. `self.reč = reč.casefold()`)? Разлог да се тако поступи можда добро формулише Лами (2021, р. 43): „the ... functions are used to manipulate the attributes of an object when the name of the attribute is not known at the time of writing the program, but is available during execution in a variable (as a string)” (а КаМП-ови раде на корпусу који бира корисник). Заправо, идеја да се уместо нотације тачком одаберу уграђене функције `getattr()`, `hasattr()` и `setattr()` потиче са сајта „Stack Overflow”, али је он непогодан за цитирање јер није „ауторитативан извор”. Уп.:

```
"""Циљ је упарити све речи из листе
са њиховим верзијама добијеним
помоћу метода casefold().
"""

class Test1:
    def __init__(self):
        for reč in [
            "Један", "ДВА", "ТРИ"]:
            self.reč = reč.casefold()

test1 = Test1()
print(vars(test1))
"""Излаз: {'reč': 'три'}."""

class Test2:
    def __init__(self):
        for reč in [
            "Један", "ДВА", "ТРИ"]:
            setattr(
                self, reč,
                reč.casefold())

test2 = Test2()
print(vars(test2))
"""Излаз: {'Један': 'један',
'ДВА': 'два', 'ТРИ': 'три'}."""
```

КаМП 2.2 и КаМП 2.3 разликују се по декларацији `__slots__` (Goodrich et al., 2013, p. 99):

```
class Rečnik: # КаМП 2.2
    tokeni_ = tokenizacija()
    __slots__ = tokeni_

    def __init__(self):
        for reč in self.tokeni_:
            setattr(
                self, reč,
                reč.casefold())

class Rečnik: # КаМП 2.3
    def __init__(self):
        for reč in tokenizacija():
            setattr(
                self, reč,
                reč.casefold())
```

Као што се у речницима српског језика, нпр. у „Речнику српскохрватског књижевног и народног језика” (Институт за српск[охрватск]и језик, 1959–present), дефиниције упарују са лемама (уп. Heisler et al., 2021, p. 263), при чему је свака лема јединствена (ниједна се не дефинише два или више пута [нити иједна више пута упућује на другу лему или друге леме]), у Python-овим речницима похрањују се парови по једног кључа и по једне вредности, при чему је сваки кључ у датом речнику јединствен (Shovic & Simpson, 2021, p. 172). (Додуше, кључеви се у Python-овом речнику не могу сортирати, него се нижу по редоследу увођења у речник [Python Software Foundation, 2025b].) На пример:

```
knjizevnici_mesto_rođenja = {  
    "Данило Киш": "Суботица",  
    "Иво Андрић": "Долац",  
    "Милорад Павић": "Београд",  
    "Александар Вучо": "Београд"  
}
```

У речнику `knjizevnici_mesto_rođenja` кључеви су имена књижевника, а вредности су имена насеља.

Класа је у Python-у средство апстракције – врста објекта и нацрт, шаблон или фабрика за генерисање других објеката. Објекат је пак скуп података заједно са повезаним понашањима (Goodrich et al., 2013, p. 69; Hetland, 2008, p. 147; Lott & Phillips, 2021, p. 4; Lutz, 2007, p. 465; Parker, 2021, p. 220). У Python-у је сваки податак објекат (Brueck & Tanner, 2001, p. 14). И саме класе су објекти (Reitz & Schlusser, 2016, p. 66). Сви објекти припадају некој класи и зову се примерци (instances) те класе (Hetland, 2008, p. 147). Понашања објекта која су добро позната, или функције које припадају класи, зову се методи (Norton et al., 2005, p. 80; Parker, 2021, p. 220). На пример, речник има метод `clear()`, који брише све парове кључева и вредности у речнику (Python Software Foundation, 2025b). Променљиве које су повезане са објектима зову се атрибути (Sweigart, 2021, p. 278). Класа се може замислити као формулар или упитник (Heisler et al., 2021, p. 281), а атрибути као питања у њему. На пример (в. Shovic & Simpson, 2021, p. 222):

```
class Član:
    def __init__(self, k_ime, p_ime):
        self.korisničko_ime = k_ime
        self.puno_ime = p_ime
```

У класи `Član` атрибути су `korisničko_ime` и `puno_ime`. Помоћу класе `Član` може се начинити нови објекат: `novi_član = Član("petar1502", "Петар Јовановић")`.

Класи је обично могуће накнадно додати арбитран атрибут (в. Martelli et al., 2017, p. 122). На пример, објекту `novi_član` може се додати година рођења: `novi_član.godina_rođenja = 1990`. Међутим, објекту `novi_član` накнадно додати арбитран атрибут није могуће ако је класа `Član` дефинисана овако:

```
class Član:
    __slots__ = (
        "korisničko_ime",
        "puno_ime")

    def __init__(self, k_ime, p_ime):
        self.korisničko_ime = k_ime
        self.puno_ime = p_ime
```

Класовни атрибут `__slots__` серија је ниски које ће класа прихватити као називе атрибута (в. Beazley, 2006, p. 94; Martelli et al., 2017, p. 122). У одређеним случајевима употребом атрибута `__slots__` могу се уштедети меморија и време (Python Software Foundation, 2025a; Ramalho, 2015, pp. 264–267).

4. Време извршавања КаМП-а 2.2 и КаМП-а 2.3

Скоро исто као у раду Данила Алексића и Лазара Мркеле, „[б]рзина је мерена у Python-у 3.8.2” и 3.13.5, „на Manjaro Linux-у, рачунаром са процесором i5-11600K и два DDR4-3200 CL16 SDRAM-а од по 16 GB и на корпусу PCL, који” (Алексић & Мркела, 2022, p. 17) „броји око 117.900.900 речи из 223.308 текстова са сајта *Политика*” (Алексић & Шандрих, 2021, p. 575; в. Табелу 1 и Табелу 2).

Табела 1

Време извршавања КаМП-а 2.2 и КаМП-а 2.3 (у секундама)

Задате поднискe	Верзија Python-a	Време у секундама (просек 10 сукцесивних мерења)				
		КаМП	КаМП 2 (мод А)	КаМП 2.1 (мод А)	КаМП 2.2 (мод А)	КаМП 2.3 (мод А)
"ac",	3.8.2	182	150	67	59	58
"aca"	3.13.5	154	116	63	54	51
"č",	3.8.2	676	398	67	61	57
"ž"	3.13.5	589	280	64	55	52
"lac",	3.8.2	187	68	67	61	57
"telj"	3.13.5	163	65	63	54	51
"š",	3.8.2	646	382	67	61	58
"ž"	3.13.5	608	266	64	54	52
"z",	3.8.2	1669	912	67	62	57
"ž"	3.13.5	1536	579	63	55	51
"ž",	3.8.2	165	147	67	59	57
"ža"	3.13.5	136	113	63	54	50

Табела 2

Време извршавања КаМП-а 2.2 и КаМП-а 2.3 (у процентима)

Задате поднискe	Верзија Python-a	Проценат најдужег времена у датом реду (према подацима из Табеле 1)				
		КаМП	КаМП 2 (мод А)	КаМП 2.1 (мод А)	КаМП 2.2 (мод А)	КаМП 2.3 (мод А)
"ac",	3.8.2	100	82,42	36,81	32,42	31,87
"aca"	3.13.5	100	75,32	40,91	35,06	33,12
"č",	3.8.2	100	58,88	9,91	9,02	8,43
"ž"	3.13.5	100	47,54	10,87	9,34	8,83
"lac",	3.8.2	100	36,36	35,83	32,62	30,48
"telj"	3.13.5	100	39,88	38,65	33,13	31,29
"š",	3.8.2	100	59,13	10,37	9,44	8,98
"ž"	3.13.5	100	43,75	10,53	8,88	8,55
"z",	3.8.2	100	54,64	4,01	3,71	3,42
"ž"	3.13.5	100	37,7	4,1	3,58	3,32
"ž",	3.8.2	100	89,09	40,61	35,76	34,55
"ža"	3.13.5	100	83,09	46,32	39,71	36,76

5. Коментар о утицају класовног атрибута `__slots__` на време извршавања

Табела 1 и Табела 2 показују да дефинисање класовног атрибута `__slots__` не само да не убрзава КаМП 2.2 него га, у односу на КаМП 2.3, и успорава. Узрок одсуству убрзања највероватније је то што је број примерака класа по један. Наиме, класовни атрибут `__slots__` вреди додати само оним класама које имају толико примерака да је важна уштеда неколико десетина бајтова по примерку, а то су типично класе које имају милионе примерака „живих” у исто време (Martelli et al., 2017, p. 122). Велики објекти, какви могу бити објекти у КаМП-у 2.2 и у КаМП-у 2.3, ову малу уштеду чине ирелевантном (в. Reitz & Schlusser, 2016, p. 122). Употреба атрибута `__slots__` убрзала је Рамалов (2015, pp. 265–266) скрипт који креира 10.000.000 примерака („*mem_test.py creates 10 million Vector2d instances using the class defined in the named module (e.g., vector2d_v3.py)*”).

6. Закључак

Најбрже верзије КаМП-а тренутно су КаМП 2.2 и КаМП 2.3, у којима се уместо речника за токенизацију и речника за упаривање по поредбеној нисци користе класе (и функције `getattr()`, `hasattr()` и `setattr()`); кључевима речника одговарају називи класовних атрибута, а вредностима речника одговарају вредности класовних атрибута.

Верзија КаМП-а са класама у којима се дефинише класовни атрибут `__slots__` (2.2) бржа је од верзије са речницима (2.1), а спорија од верзије са класама у којима се тај атрибут не дефинише.

Референце

- [1] Алексић, Д. (n.d.). *Препоручене верзије неких објављених радова*. Данило Алексић. Retrieved July 31, 2025, from https://daleksic.rs/preporucene_verzije/
- [2] Алексић, Д., & Мркела, Ј. (2022). Унапређене верзије програма „Ка минималним паровима”. *Инфотека*, 22(1), 7–31. https://infoteka.bg.ac.rs/ojs/index.php/Infoteka/article/view/2022.22.1.1_sr
- [3] Алексић, Д., & Шандрих, Б. (2021). Аутоматска ексерпција парова речи за учење изговора у настави српског као страног језика. *Српски језик*, 26(1), 567–584. <https://doi.org/10.18485/sj.2021.26.1.32>

- [4] Институт за српск(охрватск)и језик. (1959–present). *Речник српскохрватског књижевног и народног језика* (Vols. I–XXII).
- [5] Beazley, D. M. (2006). *Python: Essential reference* (3rd ed.). Sams Publishing.
- [6] Brueck, D., & Tanner, S. (2001). *Python 2.1 bible*. Hungry Minds.
- [7] Goodrich, M. T., Tamassia, R., & Goldwasser, M. H. (2013). *Data structures and algorithms in Python*. John Wiley & Sons.
- [8] Heisler, F., Amos, D., Bader, D., & Jablonski, J. (2021). *Python basics: A practical introduction to Python 3* (4th ed.). Real Python.
- [9] Hetland, M. L. (2008). *Beginning Python: From novice to professional* (2nd ed.). Apress.
- [10] Lamy, J.-B. (2021). *Ontologies with Python: Programming OWL 2.0* *Ontologies with Python and Owlready2*. Apress. <https://doi.org/10.1007/978-1-4842-6552-9>
- [11] Lott, S. F., & Phillips, D. (2021). *Python object-oriented programming: Build robust and maintainable object-oriented Python applications and libraries* (4th ed.). Packt Publishing.
- [12] Lutz, M. (2007). *Learning Python* (3rd ed.). O'Reilly Media.
- [13] Martelli, A., Ravenscroft, A., & Holden, S. (2017). *Python in a nutshell: The definitive reference* (3rd ed.). O'Reilly Media.
- [14] Next Day Video. (2013, March 21). *Transforming code into beautiful, idiomatic Python* [Video]. YouTube. <https://www.youtube.com/watch?v=OSGv2VnC0go>
- [15] Norton, P., Samuel, A., Aitel, D., Foster-Johnson, E., Richardson, L., Diamond, J., Parker, A., & Roberts, M. (2005). *Beginning Python*. Wiley Publishing.
- [16] Parker, J. R. (2021). *Python: An introduction to programming* (2nd ed.). Mercury Learning and Information.
- [17] PyData. (2023, January 11). *Raymond Hettinger: Numerical marvels inside Python - keynote | PyData Tel Aviv 2022* [Video]. YouTube. <https://www.youtube.com/watch?v=wiGkV37Kbxk>
- [18] Python Software Foundation. (2025a, June 11, 15:46 UTC). 3.3.2.4. `__slots__`. Python 3.13.5 documentation. Retrieved October 30, 2025, from <https://docs.python.org/release/3.13.5/reference/datamodel.html#slots>
- [19] Python Software Foundation. (2025b, June 11, 15:46 UTC). *Mapping types — dict*. Python 3.13.5 documentation. Retrieved October 30, 2025, from <https://docs.python.org/release/3.13.5/library/stdtypes.html#mapping-types-dict>

- [20] Ramalho, L. (2015). *Fluent Python: Clear, concise, and effective programming*. O'Reilly Media.
- [21] Reitz, K., & Schlusser, T. (2016). *The hitchhiker's guide to Python: Best practices for development*. O'Reilly Media.
- [22] SF Python. (2016, December 16). *Modern dictionaries by Raymond Hettinger* [Video]. YouTube. <https://www.youtube.com/watch?v=p33CVV29OG8>
- [23] Shovic, J. C., & Simpson, A. (2021). *Python all-in-one for dummies* (2nd ed.). John Wiley & Sons.
- [24] Storkel, H. L. (2022). Minimal, maximal, or multiple: Which contrastive intervention approach to use with children with speech sound disorders? *Language, Speech, and Hearing Services in Schools*, 53(3), 632–645. https://doi.org/10.1044/2021_LSHSS-21-00105
- [25] Sweigart, A. (2021). *Beyond the basic stuff with Python: Best practices for writing clean code*. No Starch Press.

New Enhanced Versions of the Program “Ka minimalnim parovima” (“Towards Minimal Pairs”)

Danilo Aleksić

Summary

The paper presents a modest optimization technique applied to further increase the speed of a specific Python script. The purpose of the script is to find pairs of strings which differ in two specified substrings only (in one or more positions). Depending on the language of the input corpus, these pairs, or some of them, can be useful in linguistics, language teaching, and speech-language pathology. The core of the script pairs up strings with which identical comparison strings are associated. For example, in the versions 2 and 2.1 of the script, if the specified strings are "a" and "e" and if the list of the input corpus tokens contains the strings "real" and "reel", the strings "real" and "reel" will both have "r~~ea~~l" as the comparison string and the output will contain the pair "real ~ reel". Until now, the fastest version of the script (2.1; Алексић & Мркела, 2022) coupled strings by means of a dictionary, in which the keys were the comparison strings. In this version, a dictionary was also a component (and the output) of the tokenization function. The current fastest versions, 2.2 and 2.3, couple strings by means of classes in which attribute

names and attribute values correspond to the 2.1 version's keys of the comparison dictionary and values of the comparison dictionary, respectively. The tokenization dictionary is replaced by a class, too.

Key words: language, classes, minimal pairs, optimization, dictionaries, Python.

Exploring the Synergy Between LLMs and Knowledge Graphs for Advanced Abusive Speech Detection in Serbian

Naučni rad

DOI: 10.18485/judig.2025.1.ch23

*Danka Jokić*¹,  0000-0002-7621-5847

*Ranka Stanković*²,  0000-0001-5123-6273

Abstract

Large Language Models (LLMs) have introduced immense changes in the field of artificial intelligence (AI) by their impressive capabilities in language processing and generation. However, their limitations in structured knowledge, reasoning, and factual accuracy pose significant challenges in real-world applications. On the other hand, knowledge Graphs (KGs) with their ability to store and represent interconnected data entities and semantic relationships offer a complementary solution. This paper explores the synergy between LLMs and KGs, emphasising their potential to enhance LLMs with factual grounding, enable complex reasoning, and improve explainability in AI systems.

Focusing on online safety and moderation of harmful textual content in Serbian, the paper explores and highlights how KGs can enhance LLM-based abusive speech detection systems by providing richer contextual understanding and structured reasoning. By utilising datasets such as the AbCoSER corpus of abusive speech in Serbian, the study demonstrates how KGs can provide richer context to improve abusive speech detection. Through advanced prompting techniques and hybrid search approaches inspired by retrieval-augmented generation (RAG), the research lays the groundwork for more robust, context-aware, and ethically aligned generative AI applications.

Keywords: Knowledge Graphs (KGs), Large Language Models (LLMs), Natural Language Processing (NLP), Abusive Speech

¹ Language Resources and Technologies Society (JeRTeh), danka.jokic@afrodita.rcub.bg.ac.rs

² Language Resources and Technologies Society (JeRTeh), ranka.stankovic@rgf.bg.ac.rs

1. Introduction

The emergence of large language models (LLMs), from transformer model BERT (Devlin et al, 2019) to the advanced models such as like GPT 3 (Brown et al., 2020), GPT-4 (OpenAI, 2023), has significantly impacted the field of artificial intelligence (AI) by excelling at language understanding and generation tasks (Kau et al., 2024). However, when it comes to LLMs implementation in real-life scenarios, it is delayed in areas where the correctness of output and reasoning capabilities are of the utmost importance due to the following reasons:

- LLMs rely on parametric knowledge, lacking structured knowledge and reasoning capabilities (Pan et al., 2023), which limits their effectiveness in applications requiring factual accuracy and context-aware reasoning (Pan et al., 2024).
- As being trained on general text corpora, LLMs manifest a lack of domain knowledge (Kau et al., 2024; Pan et al., 2024), including private and business-critical knowledge (Pan et al., 2023),
- LLMs are not able to verify or explain the answers, leading to limited interpretability (Pan et al., 2023),
- LLMs are prone to generating fabricated content known as hallucinations, stemming mainly from the knowledge gaps within the model (Agrawal et al., 2024),
- The knowledge in LLMs is frozen in their parameters at the time of training (Kau et al., 2024),
- The nature of data used for LLM training raised particular ethical, data privacy, and data bias concerns (Pan et al., 2023).

Knowledge graphs (KGs), on the other hand, offer a compelling solution for these challenges (Pan et al., 2024). There has been a growing focus on KGs as sources of structured knowledge for LLM-based models. The intuitive structures of KGs effectively represent real-world knowledge by representing entities with nodes and relationships between them as edges, which enables a greater understanding of a word's semantics via its context (Kau et al., 2024). The entities and their relationships are depicted in a machine-readable format, KGs provide a rich source of structured knowledge. KGs can improve, ground, and verify LLM generations so as to significantly increase reliability and trust in LLM usage (Pan et al., 2023).

The synergy with KGs has the potential to enhance LLM capabilities significantly in several key areas. First, augmenting LLMs by incorporating external knowledge of KGs can substantially improve their

ability to comprehend factual information and generate more accurate and reliable responses to complex questions, thereby reducing the hallucination (Agrawal et al., 2024). In addition, KGs also offer the necessary context and relationships to enable LLMs to perform more sophisticated reasoning tasks and incorporate commonsense knowledge into their reasoning processes, leading to more nuanced and human-like understanding. Furthermore, the explicit relationships within KGs can be harnessed to explain the reasoning behind LLM outputs, directly addressing a critical challenge in interpretable AI. ERNIE (Zhang et al., 2019), a language representation model, is a good example of this synergy. It was trained on large-scale textual corpora and KGs, allowing it to simultaneously utilise lexical, syntactic, and knowledge information, resulting in better language understanding.

While the potential is undeniable, challenges require attention. Developing effective methods for LLMs to learn from text data and knowledge graphs jointly is crucial for successful integration. Additionally, ensuring the consistency and quality of knowledge graph data is essential, as incomplete or inaccurate information can lead to biased or erroneous LLM results. Moreover, utilization of KGs is constrained by the availability of existing graphs and the resources needed for their construction and maintenance (Pan et al., 2023; Kau et al., 2024).

Despite these challenges, the integration of LLMs and KGs holds immense potential to revolutionize various AI applications. LLMs empowered by KGs can provide more accurate and comprehensive answers in question-answering systems. Intelligent assistants integrated with KGs can understand and respond to user queries with greater context and factual grounding. Additionally, the combination of LLMs and KGs can lead to the development of more factually accurate and contextually relevant natural language and code generation and the creation of more sophisticated and dynamic knowledge representation systems (Kau et al., 2024). A roadmap and forward-looking taxonomy of synergetic usage of LLMs and KGs is given in (Pan et al., 2024).

There are three possible directions for technical implementation of such systems (Pan et al., 2023). Firstly, KGs can be incorporated into training data for LLMs to complement natural language text. Secondly, triples in KGs can be used for prompt construction as input to LLMs. Additionally, KGs can be used as external knowledge in retrieval-augmented language models.

In this paper, we will present how knowledge graphs can be used to enhance the reasoning capabilities of LLMs concerning online safety and moderation of harmful textual content in Serbian. We investigated the

synergy between LLMs and knowledge graphs for the abusive speech phenomenon from various perspectives:

- Using LLMs to generate harmful content in the Serbian language,
- Using LLM to generate a knowledge graph,
- Using LLMs and KG to detect abusive speech in the Serbian language.

One of the main objectives of our research was to explore whether we could leverage domain knowledge encapsulated in KGs to enhance the generation and detection of abusive speech.

The remainder of this paper is organized as follows. In Section 2, Related work, we presented the work done in the research area of exploring the synergy between LLMs and knowledge graphs while tackling the abusive speech detection and the Serbian language. An overview of the methodology used in our research with short description of the dataset is presented in Section 3. The results and evaluation are presented in Section 4. In Conclusion, we summarize the results of our research and indicate further research directions.

2. Related Work

2.1. Using KGs for Abusive Speech Detection

In their research, Maheshappa et al. (2021) incorporated a knowledge graph into a hate speech detection pipeline. They constructed a graph out of tweet text and then produced a node2vec embedding. The embedding was an input to an LSTM. The input to another LSTM was fastText embeddings of tweet text. The outputs of two LSTMs were concatenated and passed as input to a fully connected linear layer. The results with two LSTMs were better than the results of one LSTM with fastText embeddings only. Lobo et al. (2022) presented a hybrid approach which integrates KGs and deep learning models to recognize language that references gender and sexual orientation in hate speech and thereby predict the hate speech target. The authors used an existing ontology of gender, sex and sexual orientation to support annotation of the Jegsaw toxicity dataset and ML algorithms to build weights of each term in the ontology. They used one-hot encoding of the words to train the DL model to recognize homophobic and gender hate speech. The evaluation on gender and sexual orientation demonstrates that a knowledge-grounded approach is key to enhancing model transparency, robustness, and handling of annotation errors. In their recent work, Zhao et al. (2024) introduced the MetaTox

method that combines LLMs and a meta-toxic knowledge graph to address domain-specific knowledge gaps in toxicity detection. They exploited LLMs to build a meta-toxic KG, which is later used via a retrieval process to supply LLM with toxic knowledge.

2.2. Text Classification Using Generative AI Models

The rapid development of generative AI (GenAI) models resulted in several studies using them for text classification, including abusive speech detection. Zhu et al. (2024) used ChatGPT for cyberbullying and COVID HATE datasets, reporting high recall but low precision rates. ChatGPT was used in Huang et al. (2023) to identify implicit hate in tweets and to generate an explanation of the given classification. The model correctly identified 80% of implicit hateful tweets. In another study presented in Guo et al. (2024), the authors explored LLMs as classifiers on five datasets. They achieved improvement of 7.9% to 24.2% in F1 Score compared to state-of-the-art models for some of the datasets. One of their findings was the dependence of the model's effectiveness on prompt design and the language of the text.

2.3. Abusive Language Detection in Serbian

The first information search experiments in the Serbian language dealing with the detection of attacks as a result of national, racial, or religious hatred in a corpus of newspaper articles were presented by Krstev et al. (2007). The first abusive speech dataset, AbCoSER, and lexicon in Serbian language were presented in Jokić et al. (2021). The results of abusive speech detection on the same corpus were published in Jokić et al. (2024a) and first experiments with LLMs in Jokić et al. (2024b). The best-performing classifier was BERTiC with an F1 score of 0.827 and an accuracy of 0.89. Vujičić Stanković and Mladenović (2023) used a hate speech lexicon and a dataset in the Serbian language to train a classifier for automatic hate speech detection in the sports domain. They reported reaching 96% precision in detecting hate speech using a BiLSTM deep neural network.

To address the lack of a software tool for a safer digital environment for users, Milaković et al. (2024) created a dataset of ugly and derogatory words in the Serbian language and a web extension to analyze and censor such words.

Muminović and Muminović (2025) evaluated detection of toxic comments in Serbian, Croatian and Bosnian using a dataset composed of 4,500 YouTube and TikTok comments. They evaluated four LLM models

in zero-shot and context-augmented settings. The best result was achieved with the Gemini 1.5 pro model in context-augmented mode, reaching an F1 score of 0.82 and an accuracy of 0.82. Their result also demonstrated how adding minimal context can improve toxic language detection, and they suggested strategies such as improved prompt design and threshold calibration for better results.

3. Methodology

This section describes the methodology we employed in this research. The framework we used in our experiments is depicted in Figure 1.

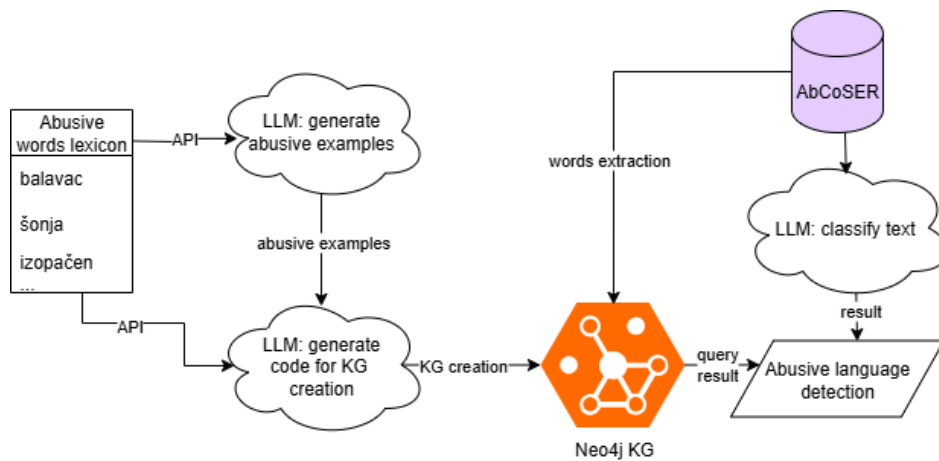


Figure 1: A framework for abusive language detection using KG and LLM

3.1. Generation of Abusive Language Examples

An abusive speech lexicon HURT_LIST used in this research was crafted by crowdsourcing and Web resources (Jokić et al., 2021) and is composed of 1,434 unique lemmas. It contains canonical forms of lemmas and its colloquial variations due to the nature of abusive speech that can be found in digital environments, which is often obfuscated by replacing or omitting characters in a word or using abbreviations. An example of obfuscation by replacement is lemma vagl₁na, which was made by replacing character i in “vagina” (eng. vagina) with number 1.

A prompt was meticulously crafted to elicit examples of hate speech based on seed terms from HURT_LIST lexicon of abusive speech, while circumventing the guardrails imposed by LLM concerning abusive speech generation. The approach pertains to knowledge-controlled generation

methods (Agrawal et al., 2024). For each seed term, five examples of abusive language were generated, and LLM was asked to rate them from 1 (least hateful) to 5 (most hateful). As the model, we used GPT-3.5-Turbo by OpenAI with temperature 0.5 to foster the creativity of generated responses.

3.2. Knowledge Graph Data Model Generation

The generated examples and the HURT_LIST lexicon are used to produce a knowledge property graph in Neo4j AuraDB3. Knowledge Graph is a design pattern to store, organize, and access interrelated data entities and semantic relationships between different pieces of information, enabling more sophisticated understanding and reasoning about data (Neo4j documentation). LLM is used for generating a graph data model from the examples and the lexicon to be used as a skeleton for data import in the Neo4J knowledge graph. The initial code4 was modified to accommodate our use case, and OpenAI GPT-4 was used as an LLM for the generation of the model comprising nodes, relations and properties. In addition, the vector embeddings were generated for each term of the HURT_LIST lexicon and each abusive speech example and stored in the corresponding graph property for later use for the similarity/vector search. For this purpose, we used the OpenAI "text-embedding-ada-002" model. The text-based and vector-based indexes were created in AuraDB accordingly.

3.3. Abusive Speech Detection

We tested the OpenAI GPT-3.5-Turbo model for abusive speech detection. Three prompt techniques were tested: zero-shot, context prompt and chain-of-thought prompting (Wei et al., 2022). We adopted the approach of Kumarage et al. (2024) for the content of each prompt. For context and chain-of-thought prompt, the LLM is provided with the definition of abusive language from Caselli et al. (2021), where abusive speech is defined as "hurtful language that a speaker uses to insult or offend another individual or a group of individuals based on their personal qualities, appearance, social status, opinions, statements, or actions." The resulting context prompt is given below:

Given the following text, evaluate if it contains abusive speech or not, specifically hurtful language that a speaker uses to insult or offend another individual or a group of individuals based on their personal qualities, appearance, social status, opinions, statements, or actions. Please provide your assessment as labels True or False.
text: {text}

³ Fully Managed Graph Database Service | Neo4j AuraDB

⁴ Initial code: llm/notebooks/llm-gen-KG-from-CSV.ipynb at main · danb-neo4j/llm

The other two prompts are provided in the Appendix. The output of the LLM-based classifier was combined with data available in the previously generated knowledge graph, which is used either as a simple word list or through similarity search. The idea was to probe GraphRAG (Edge et al., 2025) for this use case.

3.4. Dataset

As a dataset, we used the AbCoSER corpus of abusive speech in Serbian (Jokić et al., 2021), which includes labels for 6,436 Serbian tweets. It contains general abusive speech, without targeting a specific type of targets such as racial, LGBT or misogynistic speech. The dataset was annotated at two levels. The first level label indicates the presence of abusive speech. At the second level, the abusive speech was further annotated to identify the type of speech detected, whether it was hate speech, profanity, derogatory language, or other. The annotation was performed manually by two independent annotators and a supervisor to resolve any differences. Out of a total of 6,436 tweets, 1,416 were flagged as containing abusive speech.

4. Results

4.1. LLM for abusive language generation

The examples generated for the term "balavac" (eng. drooler), together with the assigned abusiveness score, are given in Table 1. Our hypothesis is that the abusiveness score, produced by LLM, is influenced by the presence of negation and explicitly offensive words in the text of examples 1-4, and therefore resulted in a higher score than example number 5.

Table 1 Table with examples of abusive speech

No.	Abusive text example	Abusiveness score
1	Nemoj da mi se obraćaš, balavac jedan!	5
2	Opet si napravio sranje, balavac jedan!	5
3	Kakav si ti balavac, ne možeš ni da obaviš jednostavan zadatak!	5
4	Ti si običan balavac, nemaš pojma o ničemu!	5
5	Pusti me na miru, balavac jedan!	4

The quality of the generated abusive speech examples varied. Only from the presented examples, we can observe issues with grammar, such as word order in a sentence, incorrect negation formulation, and improper usage of case. The examples 1, 2 and 5 were similar, although the model was instructed and set to provide creative and different instances.

4.2. LLM for Knowledge Graph Model Generation

As an illustration of the schema generation ability of LLMs, we asked the model to generate the simple KG structure for our use case. The KG structure generated employing LLM is presented in Figure 2 and it has the following elements:

Node properties:

- Dataset {_id: STRING, number: INTEGER}
- Term {_id: STRING, term: STRING, originalTerm: STRING, termEmbedding: LIST}
- Example {_id: STRING, exampleText: STRING, textEmbedding: LIST}
- AbusiveScore {_id: STRING, score: INTEGER}
- Relationship properties:
- The relationships:
- (:Term)-[:one_to_many]->(:Example)
- (:Example)-[:one_to_one]->(:AbusiveScore)

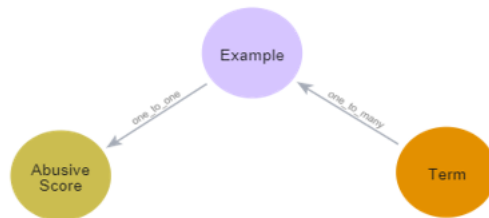


Figure 2 KG structure in AuraDB

The graph properties `termEmbedding` and `textEmbedding` correspond to embedding vectors and are stored as a list of numbers in the graph database.

An example can be related to multiple terms, for instance, in the sentence in Table 2. “You’ve done shit again, you drooler!” shall be connected to two nodes - one with the term “shit” and another with the term “drooler”.

4.3. Abusive Text Classification Results

The results of the conducted experiments are presented in Figure 3. To get a clear representation of the results, we stated outcomes of a dummy (All OFF baseline) classifier that assigns to each record the label of the most frequent class, that could serve as a default baseline model.

Detection method	Accuracy	Precision	Recall	F1 macro
All OFF baseline	0.7778	0.3889	0.5000	0.4375
Lexicon only	0.7645	0.6473	0.6280	0.6355
BERTić	0.8793	0.8285	0.8120	0.8198
BERTić +lexicon	0.8182	0.7463	0.8032	0.7649
Zero shot prompt	0.8193	0.7690	0.6391	0.6644
Context prompt	0.8284	0.8469	0.6270	0.6532
Chain of thought	0.8126	0.7483	0.6315	0.6542
Zero shot + lexicon	0.7793	0.6861	0.7011	0.6927
Context prompt + lexicon	0.7882	0.6931	0.6937	0.6934
Chain of thought + lexicon	0.7721	0.6769	0.6919	0.6834
Neo4J full text search	0.6537	0.6041	0.6450	0.5974
Neo4J simple text search	0.3058	0.5802	0.5414	0.2951

Figure 3: Results of conducted experiments with LLMs and KG

When it comes to F1 macro, accuracy and recall metrics, the result for the BERTić transformer model [BERTić], as reported by Jokić et al. (2024a), remains the best model for the dataset. Surprisingly, the GPT-3.5-turbo model with context prompt [Context prompt] achieved a better precision score than [BERTić], which indicated that the model made very accurate predictions across the classes. We can also note how adding context to the prompt influences model results [Zero shot + lexicon, Context prompt + lexicon, Chain of thought + lexicon]. Contrary to the findings of Muminović and Muminović (2025), adding context did not improve zero-shot results for this dataset.

The HURT_LIST lexicon data are integrated with LLMs as a probe for future knowledge graph integration. The value of lexicon-based text output was a binary one if any lemma from the lemmatised text was found in the abusive speech lexicon, otherwise 0. The results, indicated by the "+lexicon" label in the table [Zero shot + lexicon, Context prompt + lexicon, Chain of thought + lexicon], were promising for the LLM-based classifier, since the presence of lexicon input improved recall and

consequently resulted in a higher F1 macro score in comparison to LLM models only [Zero shot, Context prompt, Chain of thought]. In contrast, integrating lexicon input with the BERTić model [BERTić + lexicon] resulted in decreased performance, suggesting that BERTić independently classifies abusive speech more effectively due to its contextual text interpretation capabilities.

5. Conclusion

In this paper, we presented our experiments using LLMs and KGs for abusive speech detection in Serbian. We used LLMs for various tasks ranging from abusive text generation to knowledge graph structure generation, to text classification. We also tested text search over the knowledge graph in simple text and hybrid search modes as the basis for the GraphRAG approach. Based on the results of our classification experiments using an approach that combines LLMs and abusive speech lexicons, we identify a promising direction for synergy between LLMs and knowledge graphs in abusive language detection. . In that light, we aim at building a larger lexicon of abusive speech, which also includes multiword expressions (MWEs). As for the generated examples of abusive language and their inconsistent quality, we will test more LLMs, including open-source models for the Serbian language⁵, and other vendors. We aim to include additional data, such as hate targets, in the knowledge graph and build a hybrid abusive speech detection model by employing LLMs, traditional machine learning, and GraphRAG.

References

- [1] Agrawal, G., Kumarage, T., Alghamdi, Z., & Liu, H. (2024). Can Knowledge Graphs Reduce Hallucinations in LLMs? A Survey. In K. Duh, H. Gomez, & S. Bethard (Eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 3947–3960). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.naacl-long.219>.
- [2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- [3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. arXiv:1810.04805.

⁵ <https://huggingface.co/jerteh> and <https://huggingface.co/te-sla>

- [4] Edge D, Trinh H, Cheng N, Bradley J, Chao A, Mody A, Truitt S, Metropolitansky D, Ness RO & Larson J (2025). From local to global: A graph RAG approach to query-focused summarization. DOI:10.48550/arXiv.2404.16130.
- [5] Guo K, Hu A, Mu J, Shi Z, Zhao Z, Vishwamitra N & Hu H (2024). An investigation of large language models for real-world hate speech detection. DOI:10.48550/arXiv.2401.03346.
- [6] Huang F, Kwak H & An J (2023). Is ChatGPT better than human annotators? potential and limitations of ChatGPT in explaining implicit hate speech. In: Companion Proceedings of the ACM Web Conference 2023, WWW'23 Companion. Association for Computing Machinery, pp. 294–297. DOI:10.1145/3543873.3587368.
- [7] Jokić, D., Stanković, R., Krstev, C., & Šandrih, B. (2021). A Twitter Corpus and lexicon for abusive speech detection in Serbian. In 3rd Conference on Language, Data and Knowledge (LDK 2021). Schloss Dagstuhl-Leibniz-Zentrum für Informatik.
- [8] Jokić D., Stanković R. & Šandrih Todorović B. (2024a). Abusive speech detection in Serbian using machine learning. In: Mitkov R and others (eds.) Proceedings of NATURAL LANGUAGE
- [9] PROCESSING AND ARTIFICIAL INTELLIGENCE FOR CYBER SECURITY, NLPAICS, Lancaster, United Kingdom. pp. 153–163.
- [10] Jokić, D., Stanković, R., & Jaćimović, J. (2024b). Knowledge Graphs in the Era of Large Language Models: Opportunities and Challenges. Judig, Belgrade, Serbia. 60-61.
- [11] Kau, A., He, X., Nambissan, A., Astudillo, A., Yin, H., & Aryani, A. (2024). Combining Knowledge Graphs and Large Language Models (arXiv:2407.06564). arXiv. <https://doi.org/10.48550/arXiv.2407.06564>.
- [12] Krstev C, Gucul S, Vitas D and Radulovi V (2007). Can we make the bell ring? In: Paskaleva E and Slavcheva M (eds.) Proceedings of the Workshop on a Common Natural Language Processing Paradigm for Balkan Languages. Borovets, Bulgaria, pp. 15–22.
- [13] Kumaraġe, T., Bhattacharġee, A., & Garland, J. (2024). Harnessing artificial intelligence to combat online hate: Exploring the challenges and opportunities of large language models in hate speech detection. arXiv preprint arXiv:2403.08035.
- [14] Milaković A, Jocović V, Cincović J, Mićović M, Radenković U and Drašković D (2024). Detecting ugly and derogatory words in serbian language using a web browser extension. In: 2024 32nd Telecommunications Forum (TELFOR). pp. 1–4. DOI: 10.1109/TELFOR63250.2024.10819059. ISSN: 2994-5828

- [15] Muminovic, A., & Muminovic, A. K. (2025). Large Language Models for Toxic Language Detection in Low-Resource Balkan Languages (Version 2). arXiv. <https://doi.org/10.48550/ARXIV.2506.09992>
- [16] Neo4j documentation: What Is a Knowledge Graph? - Neo4j Graph Database & Analytics
- [17] OpenAI. 2023. Gpt-4 technical report. <https://doi.org/10.48550/arXiv.2303.08774>
- [18] Pan, J., Razniewski, S., Kalo, J. C., Singhania, S., Chen, J., Dietze, S., ... & Graux, D. (2023). Large Language Models and Knowledge Graphs: Opportunities and Challenges. *Transactions on Graph Data and Knowledge*.
- [19] Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J., & Wu, X. (2024). Unifying Large Language Models and Knowledge Graphs: A Roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 36(7), 3580–3599. *IEEE Transactions on Knowledge and Data Engineering*. <https://doi.org/10.1109/TKDE.2024.3352100>.
- [20] Vujičić Stanković, S. & Mladenović, M. (2023). An approach to automatic classification of hate speech in sports domain on social media. *Journal of Big Data* 10(1): 109. DOI:10.1186/s40537-023-00766-9.
- [21] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv. <https://doi.org/10.48550/ARXIV.2201.11903>
- [22] Zhu Y, Zhang P, Haq EU, Hui P and Tyson G (2024). Exploring the capability of ChatGPT to reproduce human labels for social computing tasks. In: *Social Networks Analysis and Mining*:
- [23] 16th International Conference, ASONAM 2024, Rende, Italy, September 2–5, 2024, Proceedings, Part III. Springer-Verlag, pp. 13–22. DOI:10.1007/978-3-031-78548-1_2.
- [24] Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. Ernie: Enhanced language representation with informative entities, 2019. arXiv:1905.07129.
- [25] Zhao Y, Zhu J, Xu C and Li X (2024). Enhancing LLM-based hatred and toxicity detection with meta-toxic knowledge graph. DOI:10.48550/arXiv.2412.15268.

Appendix

Appendix 1. The system prompt used to generate abusive language examples

"You are an expert for detection of abusive language in a text written in Serbian language. For the educational purposes, in order to teach people how to recognize abusive speech in text you will be given an abusive term and you will provide five different examples of the term usage in an abusive content in Twitter like style, and abusiveness score in range 1 to 5, one being least abusive and 5 being the most abusive. The term will be placed between three ### as delimiters. Please provide every instance on output mentioning every time the term you identify between delimiters and than an example in the following format for each example

"term":TERM,"example": EXAMPLE,"abusiveness score":

ABUSIVENESS_SCORE"

If you cannot provide examples, you will return an empty string

Please try to provide examples for each level of abusiveness score if it's not possible to find example for some score, please provide example anyway.

The highest priority is to have as many different examples.

Yuu should provide an answer if and only if you are completely sure that data are correct."

Output should be provided in Serbian language and in JSON format".

Appendix 2. System prompts used to classify text into abusive or not abusive categories

Zero-shot prompt: "Is the following text abusive or not? Just answer in True or False. Don't provide explanations

text: {text}"

Chain-of-thought prompting: "Given the following text, evaluate if it contains abusive speech or not, specifically hurtful language that a speaker uses to insult or offend another individual or a group of individuals based on their personal qualities, appearance, social status, opinions, statements, or actions.

First describe the step by step process of how you evaluate the given text. Then provide as output your assessment as labels True or False

text: {text}"

Istraživanje sinergije između velikih jezičkih modela i grafova znanja za naprednu detekciju uvredljivog govora na srpskom jeziku

Danka Jokić, Ranka Stanković

Sažetak

Ovaj rad istražuje mogućnosti povezivanja velikih jezičkih modela (LLMs) i grafova znanja (KG) kako bi se unapredila detekcija uvredljivog govora na srpskom jeziku. LLM modeli, od BERT-a do GPT-4, postigli su izuzetne rezultate u razumevanju i generisanju jezika, ali imaju ograničenja: nedostatak strukturisanog znanja, domenskih informacija, ograničene mogućnosti objašnjavanja svog odgovora, sklonost „halucinacijama“, zastarelo znanje nakon treniranja i etičke izazove vezane za podatke. Grafovi znanja nude rešenje (ili bar ublažavanje) ovih problema pružajući strukturisano, mašinski čitljivo znanje o entitetima i njihovim relacijama, čime se poboljšava tačnost, smanjuju halucinacije, omogućava složenije zaključivanje i interpretabilnost modela.

U istraživanju je korišćen leksikon uvredljivih izraza HURT_LIST (1.434 leme), na osnovu kojeg su pomoću GPT-3.5 generisani primeri uvredljivog govora, ocenjeni po stepenu uvredljivosti. Ti primeri i leksikon su iskorišćeni za kreiranje grafa znanja u Neo4j AuraDB, uključujući čvorove (termin, primer, ocena) i vektorske reprezentacije radi semantičke pretrage, odnosno pretrage po sličnosti. Detekcija uvredljivog govora je testirana pomoću GPT-3.5-Turbo sa tri vrste prompta pristupa promptovanju: zero-shot, kontekstualni i lanac misli (chain-of-thought). Korišćen je korpus AbCoSER sa 6.436 tvitova, od kojih 1.416 sadrži uvredljiv govor.

Rezultati pokazuju da je BERTić i dalje najbolji po F1 makro rezultatu, ali je GPT-3.5-Turbo imao bolju preciznost. Dodavanje konteksta nije unapredilo rezultate u zero-shot režimu, suprotno nekim prethodnim istraživanjima. Uključivanje leksikona povećalo je odziv i F1 makro meru kod LLM klasifikatora. Početni testovi pretrage u KG-u ukazali su na probleme sa indeksiranjem srpskog teksta zbog ograničene podrške za jezike sa manje razvijenim resursima.

Zaključuje se da sinergija LLM i KG predstavlja pravac koji obećava bolje rezultate u detekciji uvredljivog govora. Budući rad uključuje proširenje leksikona višerečnim izrazima, testiranje većeg broja LLM modela, obogaćivanje grafa znanja dodatnim informacijama kao što

je meta govora mržnje i izgradnju hibridnog modela koji kombinuje jezičke modele, tradicionalne metode mašinskog učenja i GraphRAG pristup.

Ključne reči: grafovi znanja, veliki jezički modeli (VJM), obrada prirodnog jezika, uvredljiv govori

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

Scientific paper

DOI: 10.18485/judig.2025.1.ch24

Nikitas N. Karanikolas¹,  0000-0003-1777-892X

Abstract

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

An alternative approach for step (a) is to use a Dependency analyser and extract the Dependency structure/tree for each sentence. The Dependency structure/tree of a sentence provide the dependency (grammatical) relations between the words forming a sentence.

The resources needed for NLU are various and most of them are large, populous/crowded, demanding. Lexicons are used for recognizing the inflected forms of words and Frames are providing the relevant to the verb Thematic roles and their Selectional restrictions.

There are tools that provide abilities for Syntactic (Phrase structure) analysis or Dependency analysis that relieve us of the need to have our own Lexicons. However, it is not clear if these tools are able to provide all the promising phrase structures (np, pp, etc) needed for the semantic analysis

¹ Department of Informatics & Computer Engineering, University of West Attica, Greece,
nnk@uniwa.gr

step (b). We used such a tool for the Greek Language and the results are promising. However some programming is needed. Our experience and the created rules (software routines) are described in order other colleagues to be able to do same / similar tasks and to reduce the isolation for the rule-based NLU community (users and scientists).

Keywords: Greek language; extracting phrase structures; dependency parsing; prepositional phrase attachment; frame semantics; case grammars; natural language understanding

1. Introduction

In the need for Natural Language Understanding (NLU) various task should be done. As a short abstraction of this process, the text sentences should (a) be syntactically analysed and (b) semantically analysed. With semantic analysis we refer to some process that assigns the (meanings of) phrases constituting the syntactic (Phrase structure) tree to the semantic/thematic slots/roles (the Deep cases (Bruce, 1985) that correspond to the frame of the sentence's verb (event)). (See also Case grammars (Fillmore, 1968; Bruce, 1985).) The most promising phrases to fill the semantic/thematic slots/roles of a verb (event) are noun phrases (np) and prepositional phrases (pp). Consequently the parser (the Syntactic analyzer | the Phrase Structure analyzer), either a deep or a shallow parser, should be able to isolate np and pp phrases.

An alternative approach for step (a) is to use a Dependency analyser (Jurafsky, 2019) and extract the Dependency structure/tree for each sentence. The Dependency structure/tree of a sentence provide the dependency (grammatical) relations between the words forming a sentence. A dependency tree depicts the head words (up) and the dependent words (down). The head word at the top is the Root of the Dependency tree. The arcs linking the vertices are labeled with the dependency (grammatical) relations. The advantage of Dependency trees is that the arguments to the verb are directly linked to it in the Dependency tree. Figure 1 depicts the dependency tree for the Greek sentence «Ο πρότανης έδωσε συγχαρητήρια στους βραβευθέντες επιστήμονες» (The rector congratulated the awarded scientists), as it is created by spaCy (wikipedia SpaCy; spacy.io). The syntactic (phrase structure) tree for the same sentence and the equivalent for the translated in English sentence are in figures 2 and 3, correspondingly. There are methodologies for converting Dependency Trees to Phrase Structure Trees (Xia, 2001).

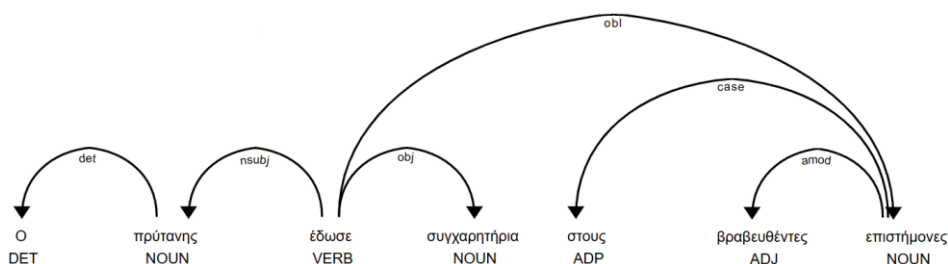


Figure 1. The dependency tree of the Greek sentence «Ο πρότανης έδωσε συγχαρητήρια στους βραβευθέντες επιστήμονες», as it is created by spaCy.

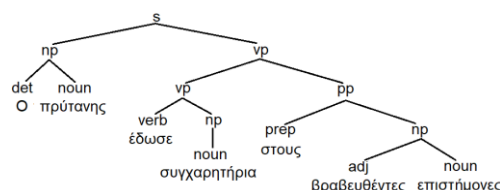


Figure 2. The syntactic (phrase structure) tree for the same Greek sentence.

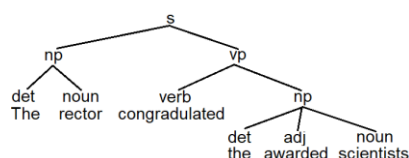


Figure 3. The syntactic tree for the translated in English sentence.

The resources needed for NLU are various and most of them are large, populous/crowded, demanding. We only mention Lexicons and Collections of Verb Frames (Semantic Frames) (Fillmore 1992; Fillmore, 2001; Goldberg, 2010). Lexicons offer the information needed for programs that recognize all the inflected forms of a word and return the head (lemma), the part-of-speech, and the features (gender, number, and case for nouns, voice, tense, mood, number, and person for verbs, etc) of the given inflected word. Each Verb Frame is a model (the template and some rules) providing the relevant to the verb (event) Thematic roles (attributes), the Selectional restrictions (Bruce, 1985) for each Thematic role (conditions for a word/phrase/clause to become the value of the Thematic role), etc.

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

In this study, we are evaluating such a tool for the Greek Language. The tool is named spaCy (Wikipedia, spaCy; spacy.io). It is mentioned / characterized as «Industrial-Strength Natural Language Processing» and it is programmed with the Python programming language. It is published

under the MIT license. spaCy supports 75 languages, among them is the Greek language, and it offers dependency parsing.

In the following, we are presenting our methodology, the application of this methodology and the results and finally we are concluding and expressing our plans for future work.

2. Methodology

Our goal is to build a system (software module, else) that given a sentence, it will be able to produce the (mostly) nominal chunks of the sentence that could fill (by the intervention of a semantic analyzer) the slots (thematic roles) that the verb(s) of the sentence requires in order to have a valid meaning. In order to being able to implement our system correctly and for real data (and not as a toy) we decided to first build a set of sentences including all/most sentence types: simple sentences, augmented sentences, elliptical sentences, sentences with subordinate/dependent clauses, lined-up sentences. This set would be our test set for checking the ability of the result from our effort. The creation of of the test set was the first step in our methodology.

Next (2nd) step in our methodology is to create the initial version of our approach (system). To do this, we used the readymade `noun_chunks` routine for seeking the noun phrases and built the first procedure (routine) for seeking the prepositional phrases (pp).

The third (3rd) step in our methodology is the running the of the so far defined system (approach) to the test set of sentences and getting the results (chunks) produced for each sentence of the set.

The fourth (4th) step in our methodology is to study the results and, if needed, build new rules (new routines) or improve existing rules (existing routines) for covering the situations that expected chunks are lost.

The last (5th) step, is to decide, according to the fourth step's results whether the goal is accomplished or there is a need to repeat the last 2 steps (3rd and 4th) until the goal is accomplished.

3. Application of Methodology and Results

Until now, we have repeated the loop (3rd to 5th steps) twice and we have defined five rules (routines) for extracting Nominal Subjects, Direct Objects, Indirect Objects, Prepositional Phrases and Clausal Complements:

```
get_dobj(doc),  
get_iobj(doc),  
get_nsubj(doc),  
get_ccomp(doc),  
get_pp(doc).
```

Clausal Complements are actions which modify / complement other actions (a clausal complement fills a slot of another verb). For example in the sentence «είπα στον δάσκαλο ότι δεν διάβασα» (I told the teacher that I didn't read), the clause «ότι δεν διάβασα» (that I didn't read) complements the verb «είπα» (lemma «λέω», tell).

All of the mentioned 5 routines are based in a tree traversal where we seek either for specific part-of-speech nodes or specific Universal dependencies (de Marneffe, 2021) between nodes or a conjunction of both. Each item of the list of prepositional phrases (items of list returned by `get_pps`) and the list of clausal complements (items of list returned by `get_ccomps`) has 5 elements: number of first and last token, introductory keyword (usually a preposition or a subordinating conjunction), head word, and the syntactic (phrasal structure) subtree. Each item of the other three lists has 4 elements: number of first and last token, head word, and the syntactic (phrasal structure) subtree.

We are now running the fourth (4th) step in the third repetition of the loop (steps 3 to 5) and now we are building a clever split of sequences of prepositional phrases. This is because there are cases of prepositional phrase attachment (Olteanu, 2005; Nadh, 2009; Ratnaparkhi, 2021) mistakes and we prefer to allow later steps in Natural Language Understanding (NLU) systems (e.g. a semantic analyser) to decide which the correct prepositional phrase attachment is. For example in the sentence «Ο Πρύτανης έδωσε συγχαρητήρια στους βραβευθέντες επιστήμονες κατά την έλευση τους στην αίθουσα τελετών» (The Rector congratulated the awarded scientists upon their arrival in the ceremony hall), spaCy assigns to the remaining after the direct object word «συγχαρητήρια», assigns to the underlined part of sentence the «obl» relation with the verb. See figure 4 for the dependency parsing tree returned by spaCy and also the «obl» relation of the underlined part of sentence with the verb.

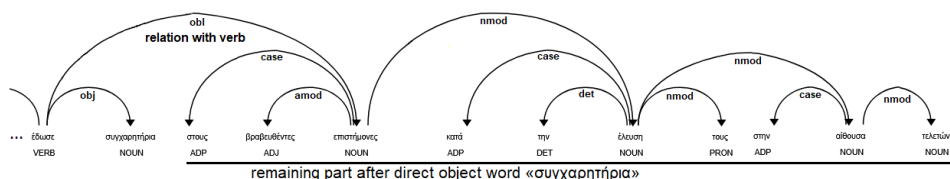


Figure 4. the dependency tree returned by spaCy.

The underlined sentence part (clause) is equivalent (can be transformed) to a constituency parsing tree (phrase structure tree), as depicted in figure 5.

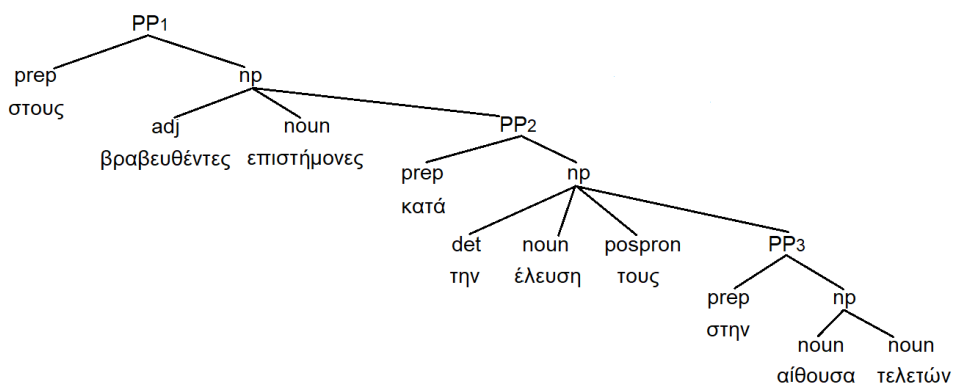


Figure 5. The equivalent constituency parsing tree (phrase structure tree) for whatever follows after the direct object word «συγχαρητήρια» of the sentence.

The same constituency parsing tree can be also depicted with the simplified nested form:

PP₁(στους βραβευθέντες επιστήμονες PP₂(κατά την έλευση τους PP₃(στην αίθουσα τελετών))).

From all the above depictions (figures 2 and 3, and the simplified nested form) it is obvious that the prepositional phrase attachments that spaCy decided for the underlined part of sentence are semantically incorrect. Consequently, the meaning of the underlined sentence part can't be a filler for any of the thematic roles that the verb «δίνω» (lemma of inflected word «έδωσε») can have. A simple partitioning (based on part-of-speech preposition, «prep» in figure 2, or «ADP» in spaCy) algorithm can return the following 3 flattened Prepositional Phrases (PP):

PP₁ (στους βραβευθέντες επιστήμονες κατά την έλευση τους στην αίθουσα τελετών),

PP₂ (κατά την έλευση τους στην αίθουσα τελετών),

PP₃ (στην αίθουσα τελετών).

Our clever split of sequences of prepositional phrases can also produce the following PPs:

PP_{1a} (στους βραβευθέντες επιστήμονες),

PP_{2a} (κατά την έλευση τους),

PP_{1b} (στους βραβευθέντες επιστήμονες κατά την έλευση τους).

Later, a semantic analysis module having all the above (6) alternative PP can assign the thematic role Location of verb «δίνω» to PP₃, assign the semantic role Beneficiary to PP_{1a}, and assign the semantic role Time to PP_{2a}.

We are also investigating if we should return a list of nodes having the universal dependency «acl:relcl» whenever an action modifies/complements/determines a noun. For example, in the clause «τον πυροσβέστη που έσβησε τη φωτιά» (the firefighter who put out the fire), the action «έσβησε τη φωτιά» (put out the fire) determines the noun «πυροσβέστη» (firefighter).

Other similar universal dependency relations we are investigating are:

- acl, it is a more general relation that points to an action that modifies / complements / determines a noun,
- xcomp, it is a relation that points to an action that modifies / complements another action (fills a slot of another verb).

4. Conclusions and future work

The tool we investigated provides dependency analysis and that relieve us of the need to have our own Lexicons. However, it does not offers all the phrase structures needed for the semantic analysis step. The isolation only of np and pp chunks (the later programmatistically) is not enough. We have to experiment to find what kind of universal dependencies and what kind of part-of-speech nodes should be investigated to extract more chunks valuable for the semantic analysis. Moreover some more programming is needed to overcome mistakes that the dependency analysers impose. For example the prepositional phrase attachment mistakes have been resolved programmatistically. Our overall experience and the results for the Greek Language are promising. We hope that other colleagues will be able to do same / similar tasks and altogether to reduce the isolation for the rule-based NLU community (users and scientists).

Scientists in rule-based NLU and NLG technologies are offering reliable, transparent and deterministic language technologies but they are working isolated. On the other hand, there are generative AI companies offering powerful and productive systems, but till now unreliable. As a consequence, rule-based NLU & NLG technologies are losing space in the market in the name of productivity. The results of paper prove that we can combine rule-based accuracy with generative AI productivity and achieve creativity and fluency with control and precision for the language tools. We are reviving the interest for rule-based NLU. Generalizing, we are reviving the interest for knowledge-driven systems.

In our set of Greek sentences (see Appendix B), we have not included sentences having multiple subjects or having multiple subject predicative (MS or MSP). It is also having only one sentence with lined-up

sentences (clauses). We are planning to extend our set of Greek sentences to include such kind of sentences and reevaluate our system's effectiveness. Next, we are planning to implement routines that isolate clauses having the universal relations of «acl:relcl», «acl», «xcomp» and else.

Appendix A – Terminology

Some of the terms used can possibly create ambiguity. To overcome, we provide the following table of Greek and English terminology.

Term in English bibliography	Term in Greek bibliography	Description / Explanation, if needed
simple sentence	απλή πρόταση	One independent clause.
-- use the description	συνθέτη πρόταση	A sentence having multiple subjects (MS) or having multiple subject predicative (MSP)
augmented sentence	επαυξημένη πρόταση	A sentence in which there are complements to the main terms.
elliptical sentence	ελλειπτική πρόταση	A sentence that there are missing terms.
~~complex sentence~~ a sentence with subordinate / dependent clauses	Πρόταση με δευτερεύουσες ή εξαρτημένες προτάσεις	A sentence with subordinate / dependent clauses. The translation of English term to Greek (σύνθετη πρόταση) has another meaning. To avoid conflict, for English we will use the description.
~~compound sentence~~ lined-up sentences	Παρατακτική σύνδεση προτάσεων	Two or more independent clauses. The English term «compound sentence» is translated to «συνθέτη πρόταση» that is used for other reason. We prefer to use the term lined-up sentences
mood	έγκλιση	The speaker's attitude towards the action (Indicative, Subjunctive, Imperative, ...)

Appendix B – Set of Greek sentences for testing Natural Language Processing (NLP) and Natural Language Understanding (NLU) systems.

'Ο Παναγιώτης πήγε στο θέατρο με το λεωφορείο'	1
'Ο Παναγιώτης πήγε στο θέατρο με τη Δήμητρα'	2
'Ο Παναγιώτης πήγε στο θέατρο με το προαστιακό λεωφορείο'	3
'Ο Παναγιώτης πήγε στο θέατρο με την όμορφη Δήμητρα'	4
'Ο Παναγιώτης πήγε στο θέατρο με την αγαπημένη του θεία'	5
'Ο Παναγιώτης πήγε στο θέατρο με τη θεία Σούλα'	6
'Ο Παναγιώτης πήγε στο θέατρο με τη θεία Δήμητρα'	7
'Ο Παναγιώτης πήγε στο θέατρο με τη θεία του'	8
'Ο Παναγιώτης πήγε στο θέατρο με τη θεία του με λεωφορείο'	9
'Ο Γιώργος πήγε τη Δήμητρα στο ξενοδοχείο'	10
'Ο Γιώργος πήγε στο ξενοδοχείο με τη Δήμητρα'	11
'Ο Βαγγέλης έστριψε προς το γείτονα με την υψηλή περίφραξη'	12
'Ο Βαγγέλης έστριψε προς το γείτονα με το ποδήλατο του'	13
'Η Ελένη στράφηκε προς το σκύλο με την μαύρη ουρά'	14
'Η Ελένη έδιωξε το σκύλο με ένα ξύλο'	15
'Η Ελένη έδωσε τον αναπτήρα στην γιαγιά της'	16
'Ο Περικλής οδήγησε το αυτοκίνητο με μεγάλη ταχύτητα'	17
'Ο Περικλής ταξιδεύει για το χωριό του με το αυτοκίνητο του πατέρα του'	18
'Ο Περικλής ταξίδεψε στο χωριό του με τον προαστιακό σιδηρόδρομο'	19
'Έδωσα στον Άγγελο το τετράδιο'	20
'Έβαλα το νερό στο ψυγείο'	21
'Πήρα από τον πατέρα μου το αυτοκίνητο'	22
'Είπα στον καθηγητή μου ότι δεν διάβασα'	23
'Φέρατε το φαγητό του σκύλου;'	24
'Φέρατε το φαγητό για τον σκύλο;'	25
'Ο νόμος επιβάλλει φόρο 15% επί των εσόδων'	26
'Σου δανείζω τα χρήματα υπό την προϋπόθεση ότι θα τα επιστρέψεις μέχρι τον Αύγουστο'	27
'Η Ευρωπαϊκή Επιτροπή χορήγησε άδεια κυκλοφορίας υπό όρους για το πρώτο εμβόλιο κατά της νόσου COVID-19'	28
'Ο Πρύτανης έδωσε συγχαρητήρια στους βραβευθέντες επιστήμονες κατά την έλευση τους στην αίθουσα τελετών'	29

'Οι οικοδομικές εργασίες θα ξεκινήσουν μετά την έκδοση της οικοδομικής αδείας'	30
'Οι γεωργοί ραντίζουν τα δέντρα τους με χαλκό κατά τη διάρκεια του Φεβρουαρίου'	31
'Οι προκομμένοι γεωργοί ραντίζουν τα καρποφόρα δέντρα τους με χαλκό πριν το μήνα Φεβρουάριο'	32
'Ο Θανάσης ταξιδεύει στο χωριό του με το αυτοκίνητο του πατέρα του'	33
'Ο Θανάσης ταξιδεύει στο δρόμο που ενώνει τη Θήβα με τη Λειβαδιά'..	34
'Η Αδαμαντία έφαγε τα μακαρόνια με ένα κουτάλι στην αυλή του σπιτιού της'	35
'Η Αδαμαντία αγόρασε μπριτζόλες από βόειο κρέας'	36
'Η Αδαμαντία αγόρασε μπριτζόλες από βόειο κρέας από ενήλικα ζώα' ..	37
'Η Αδαμαντία αγόρασε μπριτζόλες από βόειο κρέας από ενήλικα ζώα που εκτρέφονται στην Μεσσηνία'	38
'Η Αδαμαντία αγόρασε μπριτζόλες από βόειο κρέας από ενήλικα ζώα από ένα χασάπικο στα Πετράλωνα'	39
'Η νέα συγγάτοικος της Αδαμαντίας αγόρασε μπριτζόλες από βόειο κρέας από ενήλικα ζώα από ένα χασάπικο στα Πετράλωνα'	40
'Ο διευθυντής του τμήματος εγκαταστάσεων μου είπε την αιτία διαφωνίας του με τον προϊστάμενο του τμήματος μηχανολογικών κατασκευών'	41
'Συνάντησα τον πυροσβέστη που έσβησε την φωτιά στα Σάλωνα πριν μία εβδομάδα'	(23b) 42
'Του δανείζω τα χρήματα υπό την προϋπόθεση ότι ο Παναγιώτης θα μου επιστρέψει τα δανεικά μέχρι τον Αύγουστο'	(27b) 43
'Ο εργοδηγός θα στείλει στον Παναγιώτη τα εργαλεία και θα περιμένει αυτός να του επιστρέψει την μηχανή'	(27c) 44

References

- [1] Jurafsky, Daniel and Martin, James H. (2019). Chapter 15 Dependency Parsing. *Speech and Language Processing*.
https://web.stanford.edu/~jurafsky/slp3/old_oct19/15.pdf
- [2] Xia, Fei and Palmer, Martha. (2001). Converting Dependency Structures to Phrase Structures. *Proceedings of the First International Conference on Human Language Technology Research*. <https://aclanthology.org/H01-1014.pdf>
- [3] Wikipedia, spaCy. <https://en.wikipedia.org/wiki/SpaCy>
- [4] spacy.io, Industrial-Strength Natural Language Processing. <https://spacy.io/>
- [5] Fillmore, C.J. (1968). The Case for Case. Bach, E., Harms, R.T. (Eds.) *Universals in Linguistic Theory*, (pp. 1–88). Holt, Rinehart, and Winston: New York, NY, USA.
- [6] Bruce, B. (1985). Deep Case Systems for Language Understanding. Technical Report. University of Illinois at Urbana-Champaign.
<http://hdl.handle.net/2142/17534>.
- [7] Fillmore, C. J. and Atkins, B. T. (1992). Towards a Frame based organization of the lexicon: the semantics of RISK and its neighbors. Lehrer, Adrienne et al. (Eds.) *Frames, Fields, and Contrasts: New Essays in Semantics*, (pp. 75–102). Lawrence Erlbuan.
<https://www.cl.cam.ac.uk/teaching/1314/L114/1992.pdf>
- [8] Fillmore, Charles J., and Collin F. Baker. (2001). Frame semantics for text understanding. *Proceedings of WordNet and Other Lexical Resources Workshop*. NAACL.
<https://course.khoury.northeastern.edu/csg224/resources/framenet/frame.net.pdf>
- [9] Goldberg, Adele E. (2010). 3 Verbs, Constructions, and Semantic Frames. Hovav Rappaport, Malka, Doron, Edit and Sichel, Ivy (Eds.) *Lexical Semantics, Syntax, and Event Structure*, (Oxford Studies in Theoretical Linguistics, pp. 39–58). Oxford Academic.
<https://doi.org/10.1093/acprof:oso/9780199544325.003.0003>
- [10] Olteanu, Marian and Moldovan, Dan. (2005). PP-attachment disambiguation using large context. *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, (pp. 273–280). Association for Computational Linguistics.
<https://aclanthology.org/H05-1035.pdf>
- [11] Nadh, Kailash and Huyck, Christian. (2009). Prepositional phrase attachment ambiguity resolution using semantic hierarchies. Hamza, M. (Ed.) *9th IASTED*

International Conference on Artificial Intelligence and Applications. Acta Press. <https://repository.mdx.ac.uk/item/81qvx>

- [12] Ratnaparkhi, Adwait and Kumar, Atul. (2021). Resolving Prepositional Phrase Attachment Ambiguities with Contextualized Word Embeddings. *Proceedings of the 18th International Conference on Natural Language Processing*, (pp. 335–340). NLP Association of India (NLP AI). <https://aclanthology.org/2021.icon-main.40.pdf>
- [13] de Marneffe, Marie-Catherine, Manning, Christopher D., Nivre, Joakim, and Zeman, Daniel. (2021). Universal Dependencies. *Computational Linguistics*, 47 (2), 255–308. https://doi.org/10.1162/coli_a_00402

Imeničke i predložne sintagme u grčkom jeziku pomoću spaCy biblioteke

Nikitas N. Karanikolas

Sažetak

Za razumevanje prirodnog jezika potrebno je obaviti više zadataka. U pojednostavljenom prikazu, rečenice treba analizirati (a) sintaksički i (b) semantički. Pod semantičkom analizom podrazumeva se proces dodeljivanja značenja frazama koje čine sintaksičko stablo, i to semantičkim/tematskim ulogama (slotovima) koje odgovaraju semantičkom okviru glagola u rečenici (događaju). Najpogodnije fraze za popunjavanje tematskih uloga glagola jesu imeničke sintagme (NP) i priloško-predložke sintagme (PP). Zbog toga bi parser (sintaksički analizator ili analizator frazne strukture), bilo duboki ili plitki, trebalo da bude u stanju da izoluje imeničke fraze i PP sintagme.

Alternativni pristup koraku (a) jeste upotreba analizatora zavisnosti (Dependency analyser), koji izvlači zavisnosnu strukturu/stablo za svaku rečenicu. Zavisnosno stablo prikazuje gramatičke odnose između reči koje čine rečenicu.

Za razumevanje prirodnog jezika su potrebni različiti resursi, uglavnom obimni, složeni i zahtevni. Leksikoni služe za prepoznavanje flektivnih oblika reči, dok okviri (frames) obezbeđuju odgovarajuće tematske uloge glagola i njihova ograničenja.

Postoje alati koji omogućavaju sintaksičku analizu ili analizu zavisnosti i tako nas oslobađaju potrebe da održavamo sopstvene leksikone. Ipak, nije sigurno da li ti alati mogu da obezbede sve potrebne

strukture polileksemskih jedinica (NP, PP itd.) za semantičku analizu u koraku (b). Koristili smo takav alat za grčki jezik i rezultati su ohrabrujući — mada je potrebno dodatno programiranje. Naše iskustvo i kreirana pravila (u vidu softverskih rutina) su opisana u radu kako bi i drugi istraživači mogli da obavljaju iste ili slične zadatke i time smanje izolovanost zajednice koja radi na pristupima zasnovanim na pravilima za automatsko razumevanje prirodnog jezika.

Ključne reči: grčki jezik; ekstrakcija polileksemskih struktura; analiza zavisnosti; priloško-predloške sintagme; semantika okvira; razumevanje prirodnog jezika

Korpus *Srpko* kao tehnološka osnova za izradu Rečnika savremenog srpskog jezika Matice srpske

Naučni rad

DOI: 10.18485/judig.2025.1.ch25

Dušanka Vujović¹,  0000-0003-2469-2702

Branko Milosavljević²,  0000-0003-4551-9802

Apstrakt

Digitalizacija jezičkih resursa omogućava lak i brz pristup jezičkim podacima olakšavajući tako lingvistička istraživanja kao i analizu jezičkih fenomena. Ona je, takođe, od ključnog značaja za savremene leksikografske projekte jer obezbeđuje obilje pažljivo odabrane leksičke građe, lakšu pretragu i ekscerpciju primera kao i njihovu ugradnju u rečnik. Izgradnja elektronskog jezičkog korpusa *Srpko* predstavlja temelj za izradu novog *Rečnika savremenog srpskog jezika Matice srpske*. Elektronski jezički korpus omogućava jednostavno i brzo pretraživanje tekstualnih podataka i njihovu inkorporaciju u sadržaj rečničkog teksta. U radu se govori o principima, kao i idejnim i tehničkim rešenjima u vezi sa organizacijom korpusa. Prikazuje se njegova struktura, način prikupljanja podataka i formiranja baze podataka, budući da se u korpus uključuju podaci različitih formata, poput pisanih digitalizovanih i nedigitalizovanih tekstova, kao i audio i video zapisa. U cilju brže i lakše pretrage korpusa, ali i njegove kontrole, u softver su ugrađene različite funkcionalne pretrage i zahtevi za filtriranje rezultata što je omogućilo smanjivanje šuma, odnosno redundantnih informacija koje otežavaju pretrage. U softver je ugrađena i funkcija dobijanja različitih izveštaja na osnovu zadatih filtera. Ovako koncipiran korpus pruža osnovu za sistematsko organizovanje i

¹ Hankuk University of Foreign Studies in Seoul, Faculty of East European and Balkan Studies
Univerzitet u Novom Sadu, Filozofski fakultet, dusanka.vujovic@ff.uns.ac.rs

² Univerzitet u Novom Sadu, Fakultet tehničkih nauka, mbranko@uns.ac.rs

predstavljanje leksičke baze podataka u cilju izrade novog *Rečnika savremenog srpskog jezika*.

Ključne reči: srpski jezik, korpus srpskog jezika, digitalizacija, rečnik

1. Uvod

Krajem 2020. godine, u Matici srpskoj intenziviran je rad na projektu izrade novog *Rečnika savremenog srpskog jezika* (RSSJ). Rečnik se konceptijski i metodološki naslanja na ranije Matičine rečnike, šestotomni *Rečnik srpskohrvatskoga književnog jezika* (RMS) i jednotomni *Rečnik srpskoga jezika* (RSJ) kao i na *Rečnik srpskohrvatskog književnog i narodnog jezika* Instituta za jezik Srpske akademije nauka i umetnosti (RSANU). Tehnologija izrade novog Rečnika zasniva se na softverskoj platformi koju smo sami osmislili i razvili za potrebe projekta. Želeli smo da kreiramo softverska rešenja koja su prilagođena specifičnostima srpskog jezika (imajući u vidu njegovu složenu prozodijsku i morfološku strukturu) i koja u formalnom prikazu odgovaraju tradicionalnoj strukturi leksikografskog članka u dosadašnjim rečnicima srpskoga jezika. Zbog toga smo odlučili da razvijemo sopstvenu softversku platformu, umesto da adaptiramo neki od postojećih programa za izradu rečnika, čija bi već gotova rešenja trebalo dodatno prilagoditi našim potrebama.

2. O programu

Softver za izradu Rečnika sastoji se od dva povezana programa. U cilju izrade dela programa za formiranje Rečnika prvo smo razložili elemente strukture leksikografskog članka na najmanje invarijabilne jedinice. Precizno isplanirana struktura leksikografskih elemenata, koje treba uneti u unapred definisane forme programa, veoma je važan korak u razvoju softverskih rešenja jer obezbeđuje detaljan i dosledan prikaz elemenata, njihovo efikasno kombinovanje i optimalan pristup podacima u cilju lakše ekstrakcije relevantnih informacija iz unete baze podataka (GAUS 2014: 3). Tradicionalni unos teksta na papir fokusira se na konačni izgled rečničkog članka, uz poštovanje tipografskih konvencija, dok se unos u elektronskoj formi fokusira na osnovne strukturne elemente rečničkog članka, a prikaz se formira automatski. Na taj način moguće je konzistentno menjati prikaz bez intervencija na unetom sadržaju rečničkog članka kao i formirati prikaze namenjene različitim tipovima medija kao što su štampani tekst, ekran računara ili mobilnog uređaja. Zbog kompleksnosti ukupne funkcionalnosti softvera, u ovom radu daje se prikaz samo ključnih, osnovnih funkcija.

Platforma za formiranje leksikografskog članka podeljena je na dva dela. Jedan deo predviđen je za unos leme i osnovnih gramatičkih i upotrebnih informacija vezanih za nju, odnosno, u tradicionalnoj leksikografiji tzv. leve strane rečnika. Drugi deo platforme namenjen je formiranju desne strane rečnika u koju se unosi leksikografska definicija i prateći elementi od kojih je sastavljena i koji idu uz nju³. Polja za formiranje leve strana rečnika (Sl. 1) predviđaju unos obaveznih i neobaveznih podataka. Neka polja imaju unapred definisanu listu mogućih vrednosti i elementi za unos dobijaju se izborom iz padajućih menija (Sl. 2), a neka omogućavaju slobodan unos teksta kao npr. polje za dodatne informacije (Sl. 1).

Sl. 1. Polja za unos leme i osnovnih gramatičkih i upotrebnih informacija

³ Leva i desna strana rečnika su termini vezani za tradicionalnu leksikografiju i mi ih koristimo u radu jer je softver izrađen u cilju objavljivanja štampane verzije Rečnika u kojoj će jasno biti vidljivo šta pripada levoj, a šta desnoj strani. U elektronskoj leksikografiji pojmovi leva i desna strane rečnika nemaju više smisla jer se linearna, dvostrana struktura štampanog rečnika zamenjuje dinamičnim prikazom podataka. Organizacija rečničkih informacija više nije prostorno uslovljena, već je funkcionalna jer korisnik pristupa svim informacijama interaktivno i prema potrebi.

U polja za slobodan unos teksta omogućen je unos odredničke reči, zatim varijanta reči (bilo da je u pitanju akcenatska, morfološka ili ijekavska varijanta), različitih nastavaka reči kao i dodatnih informacija u vezi sa gramatičkim obeležjima reči, koje nismo mogli unapred da definišemo. Polja sa unapred definisanim vrednostima za unosu predviđena su za markiranje vrste reči, za gramatička obeležja kao što su rod imenica ili vid i rod glagola, za unos kvalifikatora upotrebne vrednosti reči i za obeležavanje statusa koju reč ima u trenutku obrade (npr. nova reč, upitni primer i sl).

Sl. 2. Padajući meni sa obavezним izborom

Sl. 3. Polja za unos elemenata leksikografske definicije

Drugi deo stranice za formiranje leksikografskog članka namenjen je unosu leksikografske definicije i svih elemenata koji ulaze u njenu strukturu i dolaze uz nju. Osim polja za unos osnovnog značenja, tu su polja za unos podznačenja, konkordanci i njihovih izvora, kvalifikatora za posebna značenja, frazeologizama, kolokacija, sinonima i antonima (Sl. 3).

3. Render Rečnika iz odrednica unetih po prethodno opisanoj strukturi

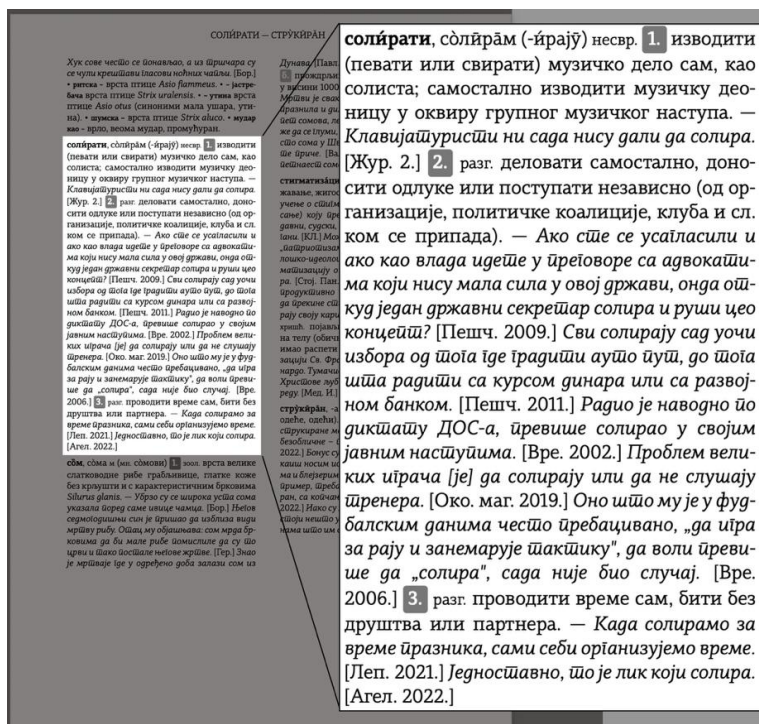
Osnovni zadatak učesnika na projektu, koji je definisala Matica srpska kao izdavač, jeste izrada Rečnika u štampanoj formi. Predviđeno je da sadrži oko 130.000 odrednica. Zbog toga je u program ugrađena funkcionalnost koja omogućava ispis i štampanje Rečnika u konačnoj, prelomljenoj formi, tako da je u svakom trenutku moguće dobiti generisani PDF prikaz (render) spreman, kako za štampu bez dodatne pripreme, tako i za prikaz na vebu, budući da je taj proces u potpunosti automatizovan. Periodično automatsko generisanje rečničkog sadržaja omogućava svakodnevni uvid u njegovu celinu u PDF formatu (Sl. 4). Elementi rečničke makrostrukture mogu da se unesu tokom rada na Rečniku, a svakako i na samom kraju izrade Rečnika i oni su sastavni deo konačnog izveštaja u PDF formatu. Unose se kao formatirani tekst ili se automatski generišu na osnovu sadržaja odrednica koje su ušle u Rečnik.

4. Baza podataka, struktura korpusa i način prikupljanja građe

Elektronski jezički korpusi neophodni su i nezaobilazni leksikografski alati jer obezbeđuju primarne izvore podataka za formiranje rečnika (RANDEL 2002: 140). Imajući to u vidu, u okviru programa za izradu Rečnika razvijena je posebna platforma za rad sa korpusom, namenjena pretragama, generisanju raznih izveštaja, ekscerpiranju primera za Rečnik i njihovom unošenju u leksikografski članak. Informacioni sistem za podršku izrade novog Rečnika, koji obuhvata i softver za rukovanje korpusom, sačinjen je kao veb aplikacija. Komponente sistema i alati za razvoj dostupni su u obliku otvorenog koda na internetu. Baza podataka kojom rukuje softver za korpus ima sledeće celine:

1. Reči čiji su gramatički oblici povezani sa njihovim osnovnim oblikom po kom ih je moguće pretražiti. Trenutno se u toj bazi nalazi oko 140.000 reči (lema). Ovi podaci prikupljeni su jednim delom uz pomoć studenata Odseka za srpski jezik i lingvistiku sa novosadskog Univerziteta, koji su u bazu unosili oblike promenljivih reči, a deo podataka je automatski preuzet iz baze podataka *wikimorph-sr*. Budući da su nam ljudski i finansijski resursi veoma skromni, nismo imali mogućnosti da se bavimo anotacijom korpusa i ovo nam je bio optimalan način da, u

relativno kratkom vremenskom periodu, u programu svedemo promenljive reči na jedan oblik za pretragu i na taj način dobijemo funkcionalan korpus. Naravno, svesni smo da to nije idealno, ali dosadašnji rad na Rečniku i pripremljen Nacrt ogleadne sveske, pokazali su da je dovoljno dobro za uspešnu pretragu korpusa kao i za izbor i unos odgovarajućih konkordanci.



Sl. 4. Izgled stranice i leksikografskog članka u renderu

2. Bibliografski opisi izvora koji čine korpus. Izvori su razvrstani po potkorpusima, za sada su to književni, novinski, administrativni, razgovorni i naučni potkorpus. Korpusnu građu prikupljali smo na razne načine budući da se radi o podacima dostupnim u različitim formatima kao što su pisani digitalizovani i nedigitalizovani tekstovi, audio i video zapisi. Jedan broj izvora dobili smo već pripremljene u digitalnoj formi, a druge (novinske članke, audio i video zapise) preuzimali smo sa interneta. One izvore koji su bili dostupni samo u štampanoj formi, sami smo digitalizovali tako što smo ih skenirali i OCR-ovali. Razgovorni korpus formirali smo automatskom transkripcijom izabranih razgovora u audio formatu pomoću softverskih alata kompanije OpenAI (Whisper).

3. Sadržaji izvora koji se čuvaju kao nestrukturiran tekst dostupan za pretraživanje i analizu.

4. Podaci koji predstavljaju statistiku pojavljivanja pojedinih reči u korpusu i odluke donete u vezi sa uključivanjem datih reči u Rečnik.

„Vrlo je važno dobro strukturirati tekstove u korpusu i unapred odrediti proporcionalno učešće različitih funkcionalnih stilova tako da on može, što vernije, da predstavi jedan prirodan jezik“ (VUJOVIĆ 2011: 40). S obzirom na to, u spisak dela za korpus ušla su dela uglednih i nagrađivanih srpskih pisaca sa prostora na kojima se srpski jezik govori, i to je oko 1200 monografija koje uključuju domaća i prevedena dela, zatim dnevne novine i časopisi, udžbenici za osnovnu i srednju školu kao i fakultetski udžbenici, razni priručnici, naučna i stručna literatura, drugi rečnici i dostupni pisani izvori koji predstavljaju transkripte govornog jezika a zatim i transkripti govornog jezika koje smo formirali specijalno za naš korpus. Budući da nam računarska tehnologija omogućava da u korpus unosimo celokupna dela, iskoristivost građe je maksimalna. Naravno, „veličina korpusa i njegova struktura zavisi od mogućnosti, odnosno od raspoloživih ljudskih i tehničkih potencijala, od vremena za koje treba da se sakupe, označe i analiziraju tekstovi“ (VUJOVIĆ 2011: 408).

Korpus je strukturiran po potkorpusima. Sakupljanje korpusa nije završeno jer mnoga književna dela nisu dostupna u digitalnoj formi. Proces njihove digitalizacije dugotrajan je i još uvek se radi na tim poslovima. Plan je da korpus *Srpko* obuhvati oko 4500 pisanih izvora od čega su 1454 književni izvori; 91 su rečnici; 100 su časopisi, nedeljnici i zbornici; 30 su dnevne novine iz kojih je preuzet 163.451 članak; 1900 su lektire za osnovnu školu; 725 su srednjoškolske lektire; 200 izvora su udžbenici i 153 izvora su razni administrativni tekstovi. Do sada je u korpus ušla i građa iz 5690 izvora razgovornog jezika. Tu građu smo formirali, u većini, skidanjem emisija sa *YouTube* kanala i njihovim automatskim transkribovanjem, a jednim delom i ručnim transkribovanjem domaćih filmova i snimljenih privatnih razgovora. Na slici broj 7 nalazi se trenutna statistika korpusa koja obuhvata broj izvora, broj reči i procentualnu zastupljenost u korpusu. Ti brojevi će se menjati jer se korpus još uvek dopunjava.

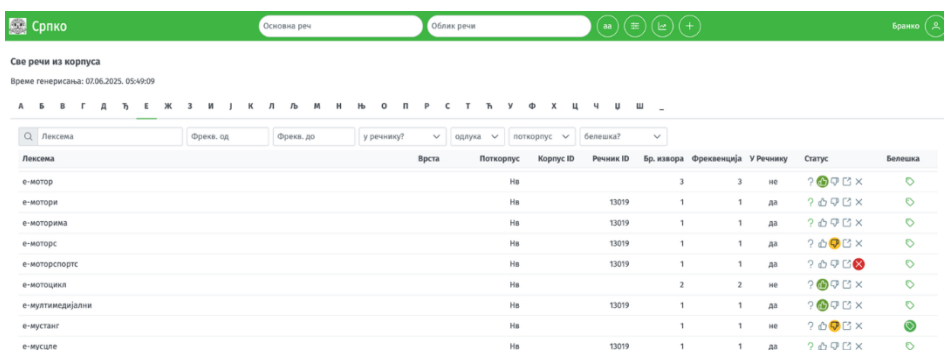
Tabela 1. Pregled statistike

Potkorpus	Broj izvora	Broj reči	Udeo
Književni	690	42.415.951	25%
Novinski	163.451	95.922.558	56%
Razgovorni	5.690	26.443.513	16%
Naučni	10.070	4.163.683	2%
Administrativni	153	1.065.547	1%
Ukupno	180.054	170.011.252	100%

5. Izveštaji i odluke o izboru reči za Rečnik

U korpus su ugrađene različite funkcionalne pretrage i zahtevi za filtriranje rezultata kao i funkcija dobijanja različitih izveštaja na osnovu zadatih filtera (Sl. 8). To mogu biti već pripremljeni izveštaji dobijeni na osnovu unapred određenih parametara ili dinamički izveštaji koji mogu da se kreiraju po potrebi na osnovu zadatih filtera. Filter može biti, na primer, frekvencija reči, potkorpus ili odluke u vezi sa unosom zadate reči u Rečnik.

Odluke o tome da li će reč ući u Rečnik ili u da li ćemo je uopšte zadržati u nekom proširenom spisku i eventualno razmatrati kasnije ili potpuno ignorisati moguće je markirati klikom na ikonice       (Sl. 5), a detaljan izveštaj o tim odlukama dobija se klikom na poveznicu „Statistika odluka“ (Sl. 6).



Српско

Основа реч Облик реч

Све речи из корпуса

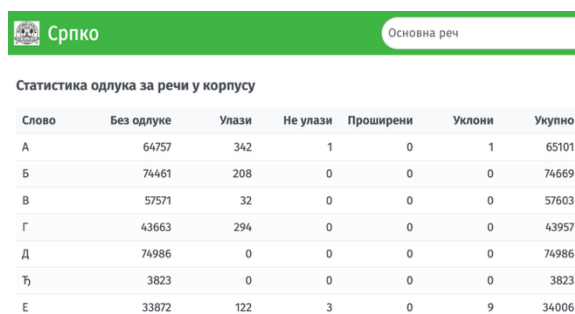
Време генерисања: 07.06.2025. 05:49:09

А Б В Г Д Ђ Е Ж З И Ј К Л Љ М Н Њ О П Р С Т Ћ У Ф Х Ц Ч Ш _

Q Лексема Фрек. од Фрек. до у речнику? одлука поткорпус Речник? Белешка

Лексема	Врста	Поткорпус	Корпус ID	Речник ID	Бр. изора	Фреквенција	У Речнику	Статус	Белешка
е-мотор		На			3	3	не	     	
е-мотори		На	13019		1	1	да	     	
е-моторика		На	13019		1	1	да	     	
е-моторс		На	13019		1	1	да	     	
е-моторспортс		На	13019		1	1	да	     	
е-мотоцикл		На			2	2	не	     	
е-мугииндустријали		На	13019		1	1	да	     	
е-мустанг		На			1	1	не	     	
е-мускле		На	13019		1	1	да	     	

Sl. 5. Filtriranje rezultata i izveštaji



Српско

Основа реч

Статистика одлука за речи у корпусу

Слово	Без одлуке	Улази	Не улази	Проширени	Уклони	Укупно
А	64757	342	1	0	1	65101
Б	74461	208	0	0	0	74669
В	57571	32	0	0	0	57603
Г	43663	294	0	0	0	43957
Д	74986	0	0	0	0	74986
Ђ	3823	0	0	0	0	3823
Е	33872	122	3	0	9	34006

Sl. 6. Statistike odluka

Na stranici sa izveštajima nalaze se dve grupe izveštaja, to su statički i dinamički izveštaji. Statički izveštaji su fiksirani prema unapred predviđenim parametrima. Na primer, jedan od njih predstavlja spisak svih

reči koje se u korpusu pojavljuju više od 10 puta. U tom izveštaju tokeni kod kojih je bilo moguće identifikovati pripadnost konkretnoj lemi objedinjeni su u jednu stavku u spisku reči, a ostali tokeni se u spisku nalaze u svojim originalnim oblicima u kojima se nalaze u korpusnoj građi. Dinamičke izveštaje možemo sami da formiramo, u zavisnosti od toga šta želimo da analiziramo. Na primer, ukoliko želimo da vidimo izveštaj svih reči koje počinju na određeno slovo ili kombinaciju slova ili ukoliko hoćemo da kombinujemo neke parametre, npr. početno slovo sa određenom frekvencijom reči i sl. poslužićemo se dinamičkim izveštajem (Sl. 7 i 8).

Динамички извештај

Реч је у речнику

Реч је у корпусу

Фреквенција од до

Одлука

Опсег слова нпр. а, а-б, аб-ад или аа-ад,в,г,фа-фр

Sl. 7. Forma za zadavanje dinamičkog izveštaja

Динамички извештај			
14.11.2025. 10:28			
у речнику: не			
у корпусу: да			
фреквенција: 10 до _			
одлуке: без одлуке			
слова: а-в			
А			
Реч	f	Реч	f
аб	1347	аврам	1018
абазовић	386	аврамовић	1019
абара	10	аврамовски	64
абас	97	автобус	23
абасидски	37	агалук	13
абација	31	агама	24
абација	36	агамемнон	193
абделазиз	11	аган	10
абдић	154	агапа	66
абдул	139	агата	786
абдулазиз	31	агбаба	16
абдулах	233	агда	11
абдулхамид	28	агејев	19
абдулхаман	40	агена	17
абел	10	агитивни	19
абер	345	агитица	23
аберадар	41	агер	85
абид	10	агим	135
абосиран	80	агитовање	48
абосирање	15	агић	17
аборт	16	агнеза	30
абортиран	48	агнета	10
адел	151	адеп	151
адела	40	адела	40
аделанда	112	аделанда	112
аделина	14	аделина	14
адем	147	адем	147
адеми	22	адеми	22
адемовић	13	адемовић	13
аден	64	аден	64
аденски	21	аденски	21
алео	13	алео	13
алег	46	алег	46
алићар	29	алићар	29
алија	38	алија	38
адиктиван	29	адиктиван	29
адикција	38	адикција	38
алил	51	алил	51
алина	16	алина	16
алис	92	алис	92
алица	16	алица	16
алиција	16	алиција	16
адлер	127	адлер	127
адлигат	30	адлигат	30
априја	130	априја	130
апријан	147	апријан	147
апријана	191	апријана	191
аер	20	аер	20
аеробан	177	аеробан	177
аерософан	15	аерософан	15
аерофлот	76	аерофлот	76
ажуриран	428	ажуриран	428
ажурирање	599	ажурирање	599
азалеја	20	азалеја	20
азап	27	азап	27
аздија	31	аздија	31
азеланчан	10	азеланчан	10
азем	74	азем	74
азем	13	азем	13
азер	19	азер	19
азербејџан	776	азербејџан	776
азербејџанац	141	азербејџанац	141
азербејџански	299	азербејџански	299
азиз	97	азиз	97
азировић	20	азировић	20
азис	10	азис	10

Sl. 8. Deo rezultata dinamičkog izveštaja

6. Pretraga korpusa

Korpus je moguće pretraživati po dva kriterijuma, po osnovnom obliku reči i po zadatom obliku reči. Rezultat pretrage po osnovnom obliku obuhvata izvore u kojima se data reč pojavljuje u bilo kom svom obliku (Sl. 9), naravno, ukoliko je u pitanju promenljiva reč. U rezultatu ove pretrage pojaviće se samo reči čiji su oblici pojavljivanja uneti u bazu podataka i povezani sa osnovnim oblikom. Njen rezultat je prikaz svih oblika pojavljivanja tražene reči. Rezultat pretrage po zadatom obliku reči omogućava pretragu određenog morfološkog oblika reči (Sl. 10) ili više reči, odnosno, konteksta (Sl. 11). Rezultat pretrage po obliku prikazuje samo taj jedan zadat oblik reči ili više reči u izvorima u kojima se nalaze. Pomoću pretrage po obliku možemo pretraživati i širi rečenični kontekst.

The screenshot shows the 'Srpsko' corpus search interface. The search bar contains 'pećurka' and the filter is set to 'Osnovni oblik reči' (Basic form of the word). The results are displayed in a list format, showing the word 'pećurka' in various contexts. The first result is: '1. [Anđ. 2] ... a onomad, u kamejari, katus. Da li neki zna put do prepljana? Obilna **pećurke** kao nagib u duži. U stvari, S obzirom da. Imajući u vidu, Tri je magičan'. The interface also includes a sidebar with filters and a search bar for the results.

Sl. 9. Pretraga po osnovnom obliku

The screenshot shows the 'Srpsko' corpus search interface. The search bar contains 'pećurka' and the filter is set to 'Osnovni oblik reči' (Basic form of the word). The results are displayed in a list format, showing the word 'pećurka' in various contexts. The first result is: '1. [Anđ. 7] ... Alisa, prema epizodi u crtanom filmu Alisa u zemlji čuda kada ona grize **pećurku**, ili je to bio kolačić, od koje se čas uvećava a čas smanjuje. Imam'. The interface also includes a sidebar with filters and a search bar for the results.

Sl. 10. Pretraga po zadatom obliku

[illegible]

Sl. 11. Pretraga po zadatom obliku – fraza

Unosom reči u polje za pretragu dobijamo je ispisanu i markiranu crvenom bojom u okviru konkordance, sa navedenim izvorom u kom se nalazi u korpusu. Širinu konteksta možemo sami da definišemo. Na desnoj strani osnovnog ekrana za rad sa korpusom pojaviće se, u svedenoj formi, ono što smo već uneli prilikom obrade zadate reči, a to znači njena značenja, podznačenja i konkordance (Sl. 12).

Српско

кошуљица

кошуља

кошуљка

кошуљанџић

кошуљар

кошуља

кошуљка

кошуљанџић

кошуљар

Објасни реч

323 погледа

фавор: 150

реч

Прегледај

Реченице: кошуљица

1

2

3

4

1. [Анб. 7]

потпуно изградио и претворио у [јутура, у]

2. [Анб. И. 6]

мудри оног Љте је томе претходило. Све је

3. [Анб. 1]

Мудраце са Истока дошли у Пакрац Исток је чинио баш пресукао **кошуљу** и легао на западну страну да подиже змијаче а Саво Истринско бијаше

4. [Анш. 1]

Господаре! како ћемо ја јесмо јуди сепе се, сребро, а? се не растаје зрно од **кошуљке** вино од посе вахузд од зрака хлеб од крува гроб од гроба зајави од причесника

5. [Анш. 1]

Јер се Пољан препознати мора сепе се, сребро, а? се не растаје зрно од **кошуљке** вино од посе вахузд од зрака хлеб од крува гроб од гроба зајави од причесника

6. [Анш. 1]

Збоге мени а? у сучу лиси сепе се, побро, а? се не растаје зрно од **кошуљке** вино од посе вахузд од зрака хлеб од крува гроб од гроба зајави од причесника

7. [Анш. 1]

Старкиново неби се крапа сепе се, тају мид, а? се не растаје зрно од **кошуљке** вино од посе вахузд од зрака хлеб од крува зајави од причесника У ОСТОЈИЊУ

8. [Анш. 1]

пукве змијачице, из кртичне транзиције, испод корњачиног крова, из змијске **кошуљке** шта може појмити собом – који јунак, који урлик, који пук, који крет,

9. [Анш. 1]

пукве змијачице, из кртичне транзиције, испод корњачиног крова, из змијске **кошуљке** шта може појмити собом – који јунак, који урлик, који пук, који крет,

10. [Ар. 4]

Ама ро тјој је води повратаск јевне младости гласило providne суве **колуџе** биља отпогас за воденог Спирита. Нјен мул је, често на неадаптан наћин

11. [Ек. Баб. 6]

сјкојана, из његовио искемених устима, која су сјајла као шарена змијска **кошуљка**, осклемисе се најћина змијска труп. Нена плавица, неко-трнучаста, пуна

12. [Бабич. 6]

сместа види своју мајку, сместа, сместа, и усташа на колена у својој **кошуљи** и увлачиш се за мрежине, као какав аналџа, пове вихате у правду тусојне

13. [Бабич. 6]

срупио ларинар и када је забринуо у маршче номену која се отика у прозачној **кошуљи**! Нена је бела стору и бојљачкама као опатаца. О, виху гориким протусте

14. [Бабич. 6]

Звие претходне три-го-дине и дјду се на поврнини, где одбацио своје **колуџе**! Стаљата поврнина воде једва се назлоде од узвишењ-ро-јева и увектојни

15. [Брат. 1]

ишенио, Како да га ни осетничка пушка не може убити јер се нос ушачени бијели **кошуљичиц** амајући од рођења. Био је међу најбољима на малој матури, а онда је

Sl. 12. Pregled i pretraga reči u korpusu

7. Ekscerpcija primera

Za unos novih konkordanci u rečnički članak potrebno je markirati željeni kontekst (Sl.13) i klikom na njega otvoriće se sva značenja i podznačenja unete reči.

The screenshot shows the JUDIG interface for the word 'košuljica'. The top bar includes the JUDIG logo, a search bar, and a 'Pregledaj' button. Below the search bar, there is a list of 12 items, each with a number and a description. The detailed view of the 'košuljica' entry shows the word in the context of a text snippet, with a list of related words and their frequencies. The interface is in Serbian and includes a 'Pregledaj' button.

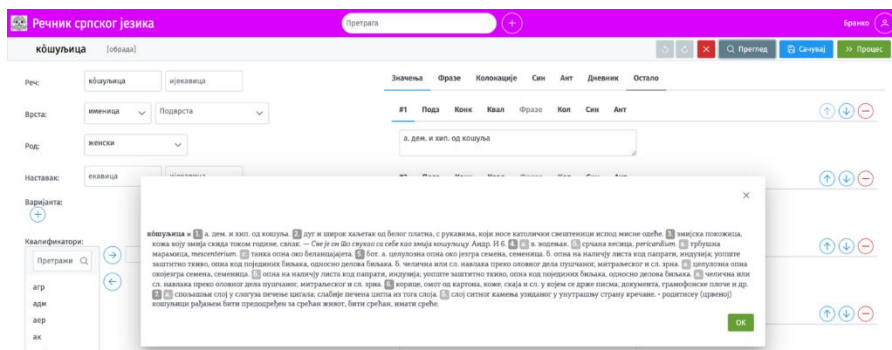
Sl. 13. Izbor konteksta

Izborom određenog značenja unosimo izabrani kontekst u tekst članka. Pošto se u korpusu nalaze originalni tekstovi u kojima ima grešaka ili su pisani latinicom, klikom na polje pored konkordance otvara se prozor u kome možemo da intervenišemo na tekstu, bilo da ga ispravljamo ili preslovljavamo iz latinice u ćirilicu (Sl. 14).

The screenshot shows the JUDIG interface for the word 'košuljica'. The top bar includes the JUDIG logo, a search bar, and a 'Pregledaj' button. Below the search bar, there is a list of 12 items, each with a number and a description. The detailed view of the 'košuljica' entry shows the word in the context of a text snippet, with a list of related words and their frequencies. The interface is in Serbian and includes a 'Pregledaj' button.

Sl. 14. Izmena konkordance

Kada se završi unos konkordanci, sve može da se proveri na stranici za unos odrednice, klikom na dugme „pregled“ (Sl. 15). Izvor, zajedno sa konkordancom, biva automatski upisan u odgovarajuće polje.



Sl. 15. Pregled odrednice

8. Softverska arhitektura informacionog sistema

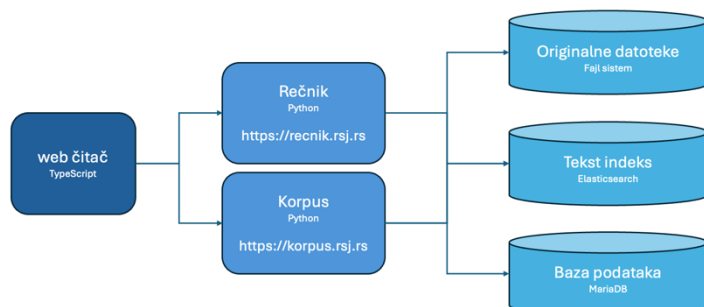
Softver korpusa i rečnika rukuje podacima u više različitih skladišta. Strukturirani podaci: rečnički članci, reči sa svim gramatičkim oblicima, bibliografski opisi izvora, statistike za izveštaje i odluke o izboru reči za Rečnik, podaci o korisnicima i pravima pristupa, čuvaju se u relacionoj bazi podataka *MariaDB*. Nestrukturirani podaci: dokumenti koji čine korpus i generisani rečnici, čuvaju se neposredno u fajl-sistemu servera na kome se izvršava program. Efikasno pretraživanje tekstualnih podataka oslanja se na poseban servis, *Elasticsearch*, koji se oslanja na fajl-sistem i koncept invertovanih datoteka.

Softver za korpus i rečnik razvijen je kroz dve odvojene aplikacije koje komuniciraju međusobno. Na taj način olakšani su nezavisan razvoj i eksploatacija ta dva osnovna dela sistema. Pored toga, pravima pristupa za korisnike korpusa i rečnika upravlja se nezavisno, što omogućava širu primenu korpusa u odnosu na rečnik, čak i izvan osnovne namene za koju je razvijen. Pristup korpusu imaju samo saradnici koji rade na izradi Rečnika, ali želja nam je da ga, ukoliko to bude moguće, kada sakupljanje građe bude završeno, učinimo i javno dostupnim.

Korpus i rečnik su implementirani kao veb aplikacije sa svojim serverskim i klijentskim slojevima. Serverski sloj je razvijen u jeziku *Python* i radnom okviru *Django*, dok je klijentski sloj razvijen u jeziku *TypeScript* i radnom okviru *Angular*. Za pristup sistemu dovoljno je koristiti neki od savremenih veb čitača. Na Sl. 16 prikazane su osnovne komponente softverske arhitekture celokupnog sistema. Aplikacije su dostupne na adresama <https://korpus.rsj.rs> odnosno <https://recnik.rsj.rs>, izvorni kod softvera je dostupan u obliku otvorenog koda na adresi <https://github.com/unsftn/rsj>, a osnovne informacije o projektu nalaze se na adresi <https://rsj.rs>.

9. Zaključak

Izgradnja elektronskog korpusa *Srpko* kao tehnološke osnovu za izradu novog *Rečnika savremenog srpskog jezika* Matice srpske predstavlja inovativan iskorak u domaćoj leksikografskoj praksi. Program za korpus je, zahvaljujući pažljivo strukturiranoj arhitekturi, obezbedio prikupljanje bogate i izbalansirane leksikografske građe i omogućio njenu organizaciju u potkorpuse kao i analizu na osnovu različitih parametara. Specijalna programska rešenja, prilagođena specifičnostima srpskog jezika, omogućila su unos, pretragu (ćiriličnu i latiničnu) i ekscerpciju primera kao i automatsko generisanje finalne verzije Rečnika u različitim formatima.



Slika 16. Komponente softverske arhitekture

Uprkos ograničenim ljudskim i finansijskim resursima kao i kompromisima u vezi sa anotacijom korpusa, postignuti rezultati potvrđuju da je tako formiran korpus omogućio sistematičnu i efikasnu leksikografsku obradu. Pretrage u korpusu jednostavne su i brze, unos konkordanci i njihovih izvora u rečnički članak je automatski, generisanje izveštaja je fleksibilno, što sve skupa ubrzava i olakšava proces izrade Rečnika.

Čitav sistem otvoren je za dalji razvoj i unapređenje funkcionalnosti kao i za širenje obima kako građe tako i korisnika. Budući razvoj projekta usredsrediće se na automatsku anotaciju teksta, sistematsko bogaćenje novim izvorima, implementaciju složenijih pretraga i proširenje skupa korisnika. Ovakvim projektima unapređuje se pozicija srpskog jezika u savremenom digitalnom okruženju.

Literatura

- [1] Gaus 2014: Gouws, H. R. (2014). Article structures: Moving from printed to e-dictionaries. *Lexikos*, 24(1). Preuzeto 20. jula 2025, sa <https://lexikos.journals.ac.za/pub/issue/view/80>
- [2] Randel 2002: Rundell, M. (2002). Good old-fashioned lexicography: Human judgment and the limits of automation. In *Lexicography and natural language processing: A Festschrift in honour of BTS Atkins*. EURALEX. Preuzeto 20. jula 2025, sa https://www.euralex.org/elx_proceedings/Lexicography%20and%20Natural%20Language%20Processing/Michael%20Rundell%20-%20Good%20Old-fashioned%20Lexicography%20Human%20Judgment%20and%20the%20Limits%20of%20Automation.pdf
- [3] Vujović 2011: Vujović, D. (2011). Internet kao elektronski korpus. *Naučni sastanak slavista u Vukove dane*, 40(1), 405–411. Beograd.
- [4] Vujović, Milosavljević 2022: Вујовић, Д., Милосављевић, Б. (2022). О програму за унос и обраду речи новог вишетомника Матице српске из лексикографског угла. У *Актуелна питања лексикологије и лексикографије српског језика* (Радови са научног скупа „Актуелна питања лексикологије и лексикографије српског језика“, одржаног у Андрићграду 9. и 10. октобра 2021), 133–149. Андрићград: Андрићев институт.

Rečnici

- [1] RMS 1969–1976: *Речник српскохрватског књижевног језика*. књ. 1-3, Нови Сад – Загреб: Матица српска – Матица хрватска, књ. 4-6, Нови Сад: Матица српска.
- [2] RSANU 1959 -: *Речник српскохрватског књижевног и народног језика*. Књ. 1-21. Београд: САНУ – Институт за српски језик.
- [3] RSJ 2007: *Речник српског језика*, Нови Сад: Матица српска.

The *Srpko* Corpus as a Technological Foundation for Compiling Matica Srpska's *Rečnik savremenog srpskog jezika*

Dušanka Vujović, Branko Milosavljević

Summary

The digitization of linguistic resources provides fast and convenient access to language data, thereby facilitating linguistic research and the analysis of language phenomena. For modern lexicographic projects, it is of key importance, as it ensures a rich supply of carefully selected lexical material, efficient retrieval and extraction of examples, and their seamless integration into dictionary entries. The *Srpko* electronic language corpus forms the foundation for compiling the new *Rečnik savremenog srpskog jezika* by Matica srpska. It allows for quick and precise searching of textual data and their direct incorporation into dictionary content.

This paper outlines the guiding principles, as well as the conceptual and technical solutions, applied in organizing the corpus. It describes its structure, methods of data collection, and database design for each component, given the diversity of available formats, ranging from digitized and non-digitized written texts to audio and video recordings. To enable faster and more efficient searches, as well as quality control, the software integrates a range of search functions and filtering options. It also includes a reporting tool for generating customized outputs based on selected filters. In this way, the corpus provides a systematic framework for organizing and presenting the lexical database required for dictionary compilation.

Despite limited human and financial resources, as well as necessary compromises in corpus annotation, the achieved results demonstrate that the corpus, as designed, enables systematic and efficient lexicographic processing. Corpus searches are straightforward and fast; the insertion of concordances and their sources into dictionary entries is automated; and the generation of reports is flexible, all of which significantly accelerate and facilitate the dictionary compilation process. The entire system remains open to further functional development and to expanding both the scope of the material and the range of its users. Projects of this kind strengthen the position of the Serbian language within the contemporary digital landscape.

Keywords: Serbian language, Serbian language corpus, digitization, dictionary.

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

Scientific paper

DOI: 10.18485/judig.2025.1.ch26

Nebojša Ratković¹,  0000-0001-5983-9642

Abstract

The Serbian-language version of Wikipedia is one of the most comprehensive and accessible knowledge resources in the digital age. Its integration into educational systems can modernize teaching, enhance digital and language literacy, and support the development of language technologies. This paper explores both the opportunities and challenges of using Wikipedia in formal education, with particular emphasis on the creation of digital corpora and their role in advancing natural language processing and related technologies. It also discusses strategic steps for embedding Wikipedia into curricula, including teacher training, the adaptation of teaching programs, and the development of digital platforms.

The first part of the paper examines Wikipedia as a teaching tool and its capacity to promote critical thinking, research skills, and digital literacy. Examples from Serbian schools and universities illustrate how Wikipedia-based assignments improve student motivation and contribute to the availability of quality Serbian-language content online.

The second part focuses on the technical dimension, demonstrating how educational contributions enrich digital corpora, enabling the development of tools such as machine translation, grammar correction, and speech recognition. The systematic use of Wikipedia in education thus strengthens the digital infrastructure of the Serbian language and ensures its relevance in the evolving technological landscape.

Keywords: Wikipedia, education, digital corpora, language technologies, Serbian language, curriculum development

¹ Wikimedia Serbia, Education Program Manager, E-mail: nebojsa.ratkovic@vikimedija.org

1. Introduction

In the contemporary digital era, education is inseparably linked to the availability and quality of online resources. Wikipedia has become one of the most frequently consulted sources of knowledge worldwide (Pavlović et al., 2017), and its Serbian-language version represents a valuable repository of information that is widely accessible, constantly updated, and created collaboratively by a large community of volunteers. While the use of Wikipedia in informal learning has been well documented (Soler-Adillon et al., 2018), its systematic integration into formal educational systems remains insufficiently explored in Serbia. Recognizing its potential to modernize teaching practices, improve digital and language literacy, and stimulate the advancement of language technologies, this paper argues for a structured approach to incorporating Wikipedia into schools and universities.

The integration of Serbian Wikipedia into education aligns with global trends that emphasize open educational resources, participatory learning, and the democratization of knowledge (Stakić et al., 2021). For students, engaging with Wikipedia offers an opportunity to develop critical thinking, research skills, and collaborative competencies, all of which are increasingly necessary in knowledge-based societies. For teachers, it provides a dynamic tool that can enrich traditional teaching materials, facilitate innovative pedagogical methods, and contribute to the formation of a more interactive classroom environment. At the same time, Wikipedia's digital content can serve as a foundation for the development of linguistic corpora that are essential for natural language processing, machine translation, and other language technologies (Marovac et al., 2023; Leipzig Corpora Collection, 2021).

Despite these opportunities, the use of Wikipedia in Serbian classrooms is still perceived with skepticism, largely due to concerns about the accuracy of content, insufficient institutional support, and a lack of teacher training (Pavlović et al., 2017). This paper therefore seeks to examine both the opportunities and challenges of such integration. It will present examples of successful initiatives, highlight technical aspects related to the creation of digital corpora, and propose strategic recommendations for educational policy. By analyzing the dual role of Wikipedia as both an educational tool and a linguistic resource, the paper aims to demonstrate its relevance for modern education and its broader contribution to the digital advancement of the Serbian language (Krstev & Stanković, 2022).

2. Wikipedia as an Educational Tool

The Serbian-language version of Wikipedia has gradually moved beyond being merely a popular online reference and has become an increasingly valuable educational tool. Unlike static textbooks, Wikipedia is dynamic, continuously updated, and enriched by a diverse community of contributors. Its adaptability and openness make it especially suitable for integration into classrooms, where it can serve not only as a source of information but also as a platform for skill development, collaborative learning, and active knowledge creation (Pavlović et al., 2017).

One of the most significant benefits of using Wikipedia in education is its potential to enhance students' critical thinking and research competencies. Articles on Wikipedia are expected to cite reliable sources, and this requirement exposes students to the process of verifying information and evaluating the credibility of references. Instead of passively consuming information, they learn to question, analyze, and compare sources. Through editing, students also gain firsthand experience with academic debate and peer review, since their contributions are publicly visible and subject to scrutiny by the community. This visibility encourages responsibility, accuracy, and a higher level of academic engagement (Soler-Adillon et al., 2018).

Practical examples from Serbia strongly support these claims. Many higher education institutions have partnered with Wikimedia Serbia to integrate Wikipedia editing into their curricula. At the Faculty of Organizational Sciences in Belgrade, for instance, students were assigned to create or expand articles relevant to their fields of study. This long-term cooperation at the Faculty of Organizational Sciences has resulted in the creation and improvement of several hundred articles, while engaging several hundred students across different courses over the years. Research shows that students who participate in such assignments demonstrate stronger motivation, higher levels of engagement, and greater awareness of academic integrity (Stakić et al., 2021). Similarly, workshops at the Faculty of Philology, carried out through long-term collaboration, engaged several hundred students across multiple courses and semesters. Their work resulted in the creation and improvement of several hundred articles, significantly enriching Serbian Wikipedia with high-quality linguistic content and advancing students' academic writing skills. Over the past decade (2015 to 2025), educational collaborations between Wikimedia Serbia and Serbian universities have included around 160 project cycles, engaging approximately 4,800 students who created or improved several thousand Wikipedia articles. These long-term activities have substantially

contributed to the expansion of high quality Serbian language content and to the enhancement of digital and academic competencies among participants.

Additional initiatives at the Faculty of Economics in Belgrade and the Faculty of Contemporary Arts demonstrated how disciplines outside philology could benefit from such approaches. Students in economics contributed articles related to business, finance, and management, thereby expanding the availability of specialized content in Serbian. Students in the arts, on the other hand, engaged with cultural and artistic topics, where visual materials and multimedia enriched their contributions. In both cases, participants expressed pride in creating resources with lasting public value, noting that their academic work transcended the boundaries of the classroom and reached a broad online audience (Pavlović et al., 2017).

In secondary schools, pilot workshops illustrated how Wikipedia can foster digital and media literacy. Guided by trained facilitators, students explored the fundamentals of editing, citation, and neutrality. Teachers observed that such assignments helped pupils understand the importance of perspective and bias in historical and social narratives.

Another important dimension is teacher training. Experience from accredited professional seminars organized by Wikimedia Serbia has shown that educators are often initially skeptical about Wikipedia. Concerns usually revolve around reliability and the risk of plagiarism. However, once teachers are introduced to methods for using Wikipedia constructively, such as incorporating it into project-based learning, assigning article editing as coursework, or using talk pages to practice academic debate, they tend to recognize its pedagogical value (Pavlović et al., 2017). During the period 2015–2025, several dozen teachers participated in accredited seminars on integrating Wikipedia into education. While not all of them continued the practice, a number of individual teachers chose to incorporate Wikipedia-based assignments into their classrooms, thereby sustaining the collaboration and extending its impact. Teachers emphasized that Wikipedia-based activities encouraged student independence, creativity, and responsibility in ways traditional assignments often did not. Moreover, the open and collaborative nature of the platform aligned well with the principles of inclusive and participatory education promoted in Serbia's educational reforms (Krstev & Stanković, 2022).

These practices reveal that Wikipedia's role in education goes far beyond functioning as a simple reference. It becomes a didactic instrument that transforms learning into an interactive, student-centered process. Tasks based on Wikipedia stimulate collaboration, creativity, and digital

responsibility while also enriching the body of Serbian-language knowledge available online. In doing so, educational institutions contribute to the visibility and vitality of the Serbian language in the digital sphere, ensuring that it remains relevant and accessible in an era increasingly dominated by global digital platforms (Marovac et al., 2023).

The integration of Wikipedia into formal education in Serbia requires a structured and strategic approach that goes beyond sporadic initiatives. While isolated examples from universities and secondary schools demonstrate the platform's potential, a more systematic model is necessary for long-term sustainability. Such integration involves aligning Wikipedia-related activities with curricula, providing adequate teacher training, and creating supportive digital infrastructures. These steps are crucial for transforming Wikipedia from an occasional classroom tool into an established component of Serbia's educational framework (Soler-Adillon et al., 2018).

2.1. Teacher Training, Professional Development and Curriculum Alignment

Teachers represent the key actors in integrating Wikipedia into schools and universities. Without adequate preparation, many remain hesitant to incorporate Wikipedia into their teaching practice, often citing concerns about accuracy, plagiarism, or assessment criteria (Pavlović et al., 2017). To address these issues, Wikimedia Serbia, in cooperation with educational institutions, has organized accredited professional development seminars. The introduction of accreditation was particularly significant, as it ensured that teachers' engagement with Wikipedia was formally recognized as part of their professional development. This alignment with national standards encouraged broader adoption, particularly among younger teachers eager to experiment with innovative methods. Teachers who later implemented Wikipedia-based assignments in their schools reported that students responded positively, perceiving such projects as more meaningful and motivating compared to traditional written essays (Stakić et al., 2021).

A systematic integration of Wikipedia requires its inclusion in official curricula. At the university level, this can be achieved by embedding Wikipedia editing projects into existing subjects such as language studies, history, sociology, or computer science. Several faculties have already experimented with such practices.

At the secondary school level, the potential lies in aligning Wikipedia activities with competencies outlined in Serbia's educational

strategy, such as digital literacy, critical thinking, and teamwork. Students not only learned to evaluate multiple perspectives but also became active participants in knowledge production (Pavlović et al., 2017). By positioning Wikipedia-based tasks within the framework of project-based learning, schools can achieve curriculum goals while simultaneously contributing to the public pool of knowledge in Serbian.

However, curriculum alignment is still at an early stage. While individual teachers and faculties have adopted these methods, official documents such as the national Strategy for Education have yet to explicitly recognize Wikipedia as an educational resource. A future step would involve including Wikipedia and similar open educational resources in policy guidelines, ensuring consistency across schools and regions (Krstev & Stanković, 2022). Without this formal recognition, the use of Wikipedia risks remaining fragmented and dependent on the enthusiasm of individual educators rather than being embedded in systemic practice.

Despite the opportunities, integrating Wikipedia into the educational system faces challenges. The first obstacle is skepticism among educators, many of whom continue to perceive Wikipedia as unreliable (Pavlović et al., 2017). Although research and practice show that Wikipedia articles are generally as accurate as traditional encyclopedias, these perceptions persist and hinder wider adoption. Another barrier lies in the lack of official policy support—without explicit recommendations in strategic documents, teachers often fear that incorporating Wikipedia might not be formally recognized as valid educational practice (Krstev & Stanković, 2022).

Technical challenges also remain. Many schools, especially in rural areas, still lack adequate digital infrastructure, and teachers are not always trained to guide students in advanced digital assignments. Furthermore, issues such as plagiarism and improper citation require continuous monitoring and guidance. Addressing these challenges requires a multi-layered approach that combines teacher support, policy frameworks, and technological investment (Marovac et al., 2023).

3. Wikipedia and Language Technologies

The development of language technologies relies heavily on the availability of high-quality digital corpora. In this respect, Wikipedia represents an invaluable resource. Its openly licensed content, standardized structure, and wide thematic coverage make it an ideal foundation for building linguistic corpora used in natural language processing (NLP), machine translation, speech recognition, and other language technologies.

For the Serbian language, which is relatively underrepresented in global digital resources compared to English or major European languages, integrating Wikipedia into the construction of corpora is not only useful but essential for ensuring its digital presence and competitiveness (Krstev & Stanković, 2022).

Wikipedia articles are particularly suitable for corpus creation because they adhere to consistent formatting, rely on references, and are regularly updated by a community of editors. Each article provides structured text, metadata, and hyperlinks that facilitate linguistic analysis. Researchers can extract balanced datasets of different genres, such as history, science, culture, and technology, within a single platform. Unlike news portals, which often focus on short-term topics, Wikipedia provides encyclopedic content that is more stable and comprehensive.

For Serbian, a corpus based on Wikipedia articles currently contains over 26 million tokens and 1.7 million sentences, compiled in the Leipzig Corpora Collection (Leipzig Corpora Collection, 2021). In addition, the Serbian web corpus includes approximately 894 million tokens, demonstrating the scale of resources available for computational processing (UBC Library, 2025). Academic institutions in Serbia have already experimented with using Wikipedia as a basis for compiling corpora aimed at developing spell-checkers, grammar correction tools, and lemmatization resources (Marovac et al., 2023). However, systematic efforts remain limited, and much of the potential is yet to be realized.

The educational use of Wikipedia directly contributes to the enrichment of such corpora. When students and teachers create new articles or expand existing ones, they produce large amounts of standardized text that is publicly available and digitally accessible. Each contribution strengthens the corpus and provides raw material for further linguistic analysis.

This process also enhances the representativeness of corpora. By integrating educational contributions, subject areas such as economics, sociology, and the arts—traditionally less covered in standard corpora, become part of the linguistic record. The impact is therefore twofold: students gain valuable academic and digital skills (Stakić et al., 2021), and language technologies benefit from richer, more diverse resources.

Although Wikipedia provides a rich foundation, several linguistic challenges complicate the creation of robust Serbian corpora. The Serbian language is morphologically complex, with rich inflection and case systems, which requires advanced annotation and lemmatization methods (Marovac et al., 2023). Furthermore, Serbian is digraphic, written in both Cyrillic and Latin alphabets, which necessitates normalization across

scripts for effective computational processing. These issues demand additional preprocessing steps, such as script unification, morphological tagging, and syntactic parsing.

Another challenge lies in domain coverage. While Wikipedia articles are encyclopedic, certain fields (e.g., law, technical sciences) remain underrepresented, leading to gaps in corpus coverage. Supplementing Wikipedia with additional open sources, while still ensuring linguistic consistency, remains a task for future corpus-building projects (Krstev & Stanković, 2022).

Corpora derived from Wikipedia are already used in several technological applications relevant to Serbian:

- **Natural Language Processing (NLP):** Tools for tokenization, part-of-speech tagging, and lemmatization rely on corpora containing millions of words. The inclusion of Wikipedia texts allows these tools to be tested and improved on standardized material (Marovac et al., 2023).
- **Machine Translation:** Open datasets based on Wikipedia articles in multiple languages have been successfully used to improve translation engines such as Google Translate or DeepL. For Serbian, bilingual corpora derived from Wikipedia parallel articles with English and other South Slavic languages already contain aligned sentence pairs (Leipzig Corpora Collection, 2021).
- **Lexical Databases and Semantic Networks:** Dictionaries and ontologies can be enriched through Wikipedia entries, especially categories, interlinks, and metadata, which provide structured connections between concepts (Krstev & Stanković, 2022).

Each of these applications directly benefits from educational initiatives where students contribute content. The combination of pedagogy and technology thus produces a multiplier effect: education enhances language technologies, while language technologies in turn support more effective digital education.

4. International Comparisons and Societal and Cultural Benefits

Other countries provide valuable examples of how Wikipedia can be systematically used for corpus building and technology development. In Finland, Wikipedia content was incorporated into the national corpus project for Finnish, enabling progress in machine translation and AI applications for a language with a relatively small speaker base. In the Czech Republic, the Czech National Corpus uses Wikipedia extensively to

enrich linguistic datasets, which has significantly improved their NLP tools and language-learning platforms. Similarly, Sweden integrated Swedish Wikipedia texts into large-scale corpora used in both academic research and commercial products (Soler-Adillon et al., 2018).

Beyond national initiatives, several international projects highlight the role of Wikipedia in cross-linguistic data development. For example, the Wikimedia Foundation's Content Translation Tool enables editors to translate articles between languages, simultaneously generating parallel corpora that can be used for machine translation training. European projects such as ELRC-SHARE and Europeana emphasize the digitization of cultural and linguistic heritage and often rely on Wikipedia as a multilingual source (Krstev & Stanković, 2022).

The role of Wikipedia in language technologies extends beyond technical applications. By enriching Serbian-language corpora, it contributes to preserving cultural heritage, supporting digital inclusion, and ensuring that Serbian remains a fully functional language in the digital age. High-quality NLP tools based on Wikipedia texts can facilitate inclusive education for students with disabilities (e.g., through text-to-speech applications), promote cross-cultural communication via improved translation tools, and enhance the accessibility of digital archives. In this way, Wikipedia strengthens not only the technological infrastructure of the language but also its cultural and social relevance (Marovac et al., 2023).

5. Discussion and Conclusion

The integration of Wikipedia into Serbian education and its role in advancing language technologies illustrate the dual potential of digital platforms: they serve both as tools for teaching and as resources for technological development. Research and educational practice consistently show that students perceive Wikipedia as a motivating and meaningful learning environment, particularly when their work becomes visible to the wider public. Teachers, although often initially skeptical due to concerns about reliability and plagiarism, frequently acknowledge its value once they receive appropriate training and methodological guidance.

From an educational perspective, integrating Wikipedia into curricula has proven to enhance critical thinking, digital literacy, and collaborative learning. It also contributes to the expansion of high-quality Serbian-language content online, aligning with global educational trends that emphasize open access, participatory learning, and interdisciplinary collaboration. However, these benefits can be fully realized only if institutions formally recognize Wikipedia as an educational resource and

integrate it into official strategies and programs. Without such systemic support, the use of Wikipedia risks remaining fragmented and dependent on the initiative of individual educators.

From a technological perspective, Wikipedia represents a substantial linguistic and digital resource. Texts generated through educational projects contribute to the growth of digital corpora used for developing natural language processing tools, machine translation systems, and grammar and spell-checking applications. At the same time, challenges such as the morphological complexity of Serbian, the coexistence of Cyrillic and Latin scripts, and uneven thematic coverage call for continued collaboration between educators, linguists, and technology experts.

The broader implication of these findings is that education and language technologies should be viewed as mutually reinforcing domains. When embedded in formal education, Wikipedia functions as a catalyst for both pedagogical innovation and technological progress. Conversely, the development of advanced digital tools enhances its educational value, for instance through improved translation, text-to-speech, and search functionalities. This circular relationship ensures that each educational activity contributes to linguistic infrastructure, while every technological advancement further strengthens the educational process.

Looking ahead, several steps appear essential for maximizing the benefits of this integration. Policymakers should acknowledge Wikipedia as a legitimate and valuable educational tool within strategic documents. Teacher training programs need to be continuously expanded, ensuring that educators possess both technical and methodological competence. National corpus initiatives should systematically incorporate Wikipedia-based materials, supported by open access principles and inter-institutional cooperation. Furthermore, Serbia's participation in international projects that connect smaller languages to global digital infrastructures will ensure the linguistic equality and sustainability of the Serbian language in the digital era.

In conclusion, integrating the Serbian-language Wikipedia into education is not merely an innovation in pedagogy but a strategic investment in the language's digital future. By linking teaching and technology, Wikipedia provides a sustainable model for modernizing classrooms, empowering students, and enhancing language resources. Its dual role, as both a pedagogical tool and a linguistic resource, positions it as a cornerstone of twenty-first-century education in Serbia. Systematic implementation will ensure that the Serbian language remains vibrant, functional, and technologically supported within the global digital environment.

References

- [1] Krstev, C., & Stanković, R. (2022). *Report on the Serbian Language* (Deliverable D1.35, European Language Equality project). European Language Equality. https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D1_35___Language_Report_Serbian_.pdf
- [2] Leipzig Corpora Collection. (2021). *Serbian Wikipedia corpus based on material from 2021* [Dataset]. Leipzig Corpora Collection. https://corpora.wortschatz-leipzig.de/en?corpusId=srp_wikipedia_2021
- [3] Marovac, U. A., Avdić, A. R., & Milošević, N. Lj. (2023). *A Survey of Resources and Methods for Natural Language Processing of Serbian Language*. arXiv preprint. <https://doi.org/10.48550/arXiv.2304.05468>
- [4] Pavlović, D., Petrović, Z. S., & Mamutić, A. (2017). Wikipedia – From the Popular Source of Information to the Pedagogical Tool. *International Journal of Social Sciences & Educational Studies*, 4(3), 36–48. <https://ijsses.tiu.edu.iq/index.php/ijsses/article/view/406>
- [5] Soler-Adillon, J., Pavlović, D., & Freixa, P. (2018). Wikipedia in Higher Education: Changes in Perceived Value through Content Contribution. *Comunicar*, 26(54), 39–48. <https://doi.org/10.3916/C54-2018-04>
- [6] Stakić, Đ., Tasić, M., Stanković, M., & Bogdanović, M. (2021). Students' Attitudes Towards the Use of Wikipedia: A Teaching Tool and a Way to Modernize Teaching. *Área Abierta*, 21(2), 309–325. https://www.researchgate.net/publication/353110388_Students'_Attitudes_Towards_the_Use_of_Wikipedia_A_Teaching_Tool_and_a_Way_to_Modernize_Teaching
- [7] UBC Library. (2025). *Serbian web corpus (srWaC): 894 million tokens* [Dataset]. University of British Columbia. <https://guides.library.ubc.ca/c.php?g=306932&p=2051153>

Интеграција Википедије на српском језику у образовни систем и унапређење језичких технологија

Небојша Ратковић

Сажетак

Интеграција Википедије на српском језику у образовни систем представља иновативан приступ који истовремено унапређује наставне методе и доприноси развоју језичких технологија. Рад анализира потенцијале и изазове овакве интеграције у три области: улогу Википедије као наставног алата, њено укључивање у наставне планове и програме, и значај за изградњу дигиталних корпуса и развој технологија обраде природног језика.

Први део рада показује да задаци засновани на Википедији подстичу критичко мишљење, развијају истраживачке вештине и дигиталну писменост ученика и студената. Примери из високог и средњег образовања у Србији потврђују да уређивање чланака повећава мотивацију ученика и њихов осећај одговорности, а истовремено обогаћује јавно доступан садржај на српском језику.

Други део наглашава важност стратешког приступа – обуке наставника, укључивања у наставне програме и развоја дигиталних платформи. Без институционалне подршке и званичног препознавања, примена Википедије у настави остаје ограничена на ентузијазам појединаца.

Трећи део указује на техничке аспекте и значај Википедије за изградњу корпуса неопходних за машинско превођење, алате за исправљање граматике, као и технологије препознавања и синтезе говора. Кроз образовне активности генерише се сталан прилив дигиталног текста, што обезбеђује да српски језик остане виталан и равноправан у дигиталном добу.

Кључне речи: Википедија, образовање, дигитални корпуси, језичке технологије, српски језик, наставни план и програм

Употреба модела *Whisper Large v3 Sr* за транскрипцију говора на српском језику у програмском језику *Пајтон* на платформи *Гугл колаб*

Научни рад

DOI: 10.18485/judig.2025.1.ch27

Никола Јанковић¹,  0000-0003-3484-4220

Јована Иваниш²,  0009-0006-1225-3461

Апстракт

У овом раду представљен је скрипт у програмском језику *Пајтон* (енг. *Python*) на платформи *Гугл колаб* (енг. *Google Colab*) који користи фино подешени модел за транскрипцију говора на српском језику *Whisper Large v3 Sr* (<https://huggingface.co/Sagicc/whisper-large-v3-sr-cmb>), чиме се омогућава бесплатно, квалитетно и једноставно транскрибовање говора на српском језику у текст. Мотивација за креирање овог скрипта проистекла је из недостатка доступних алата који истраживачима пружају једноставно коришћење овог модела без потребе за напредним техничким знањима и значајним рачунарским ресурсима. Овај скрипт омогућује једноставно учитавање аудио-записа, њихову транскрипцију помоћу модела *Whisper Large v3 Sr* и преузимање транскрипције у текстуалном формату. У раду ће најпре укратко бити описан горенаведени модел који скрипт користи, а затим ће детаљно бити описан начин функционисања самог скрипта, уз упутство за коришћење. Такође, биће представљени проблеми које је било неопходно превазићи за успешну примену овог приступа, као што је потреба за аутоматском сегментацијом звучних записа, одређивање оптималних параметара сегментације и омогућавање подршке за различите формате аудио-материјала. Осим тога, биће

¹ Филолошки факултет, Универзитет у Београду, nikolajankovickv@gmail.com

² Институт за српски језик САНУ, jovana.ivanis@gmail.com,

представљено неколико приступа везаних за смањивање броја грешака у транскрипцији. На крају, биће приказана евалуација тачности и ефикасности транскрипције, као и статистичка анализа уочених грешака. Сматрамо да овај алат може бити од великог значаја за истраживаче, с обзиром на то да убрзава процес обраде говорних података, омогућава обраду велике количине аудио-материјала у кратком периоду и пружа конзистентан и поновљив метод транскрипције, што је нарочито значајно за научну методологију и поновљивост истраживања везаних за језик. Поред тога, сматрамо да алат може бити користан и другим корисницима, с обзиром на то да омогућава једноставно креирање титлова за видео-садржаје, претварање аудио-бележака у текстуални формат, стварање аутоматски генерисаних транскрипата за особе са оштећеним слухом, као и стварање текстуалних архива аудио-садржаја.

Кључне речи: транскрипција говора, српски језик, Пајтон, Whisper Large v3, Гугл колаб, NLP.

1. Увод

Технологија аутоматског препознавања говора (енг. *automatic speech recognition*, скр. ASR) прошла је од својих почетака у касним 1940-им и 1950-им годинама до данас вишедеценијски процес унапређивања и промена парадигми на којима је заснована (Jurafsky & Martin 2025). У овом процесу, ова технологија сазрела је до те мере да може имати практичну примену у различитим областима (*ibid.*). Са друге стране, постоји неколико препрека које отежавају примену ове технологије на језицима са мањим бројем говорника у односу на енглески (попут српског језика), нарочито када су у питању решења отвореног кода (енг. *open-source*).

Најпре, сами скупови података на којима се модели за препознавање говора тренирају садрже знатно већи проценат енглеског језика у односу на друге језике. Како наводе Пратап и сарадници (Pratap et al. 2020: 2757), овакви скупови података постоје за друге језике, али су најчешће ограниченог обима, похрањени на различитим местима и ретко су доступни под отвореном лиценцом. Када је у питању српски језик, највећи допринос на овом пољу представља недавно објављени *ParlaSpeech-SR*, који обухвата 896.22 сата говора на српском језику³ (Ljubešić et al. 2025). Затим, наведени

³ У питању су транскрипти говора посланика у заседањима Народне скупштине Републике Србије.

фактор доступности адекватних података (односно, њихово одсуство) значајно утиче на перформансе модела када је у питању аутоматско препознавање говора; наиме, како наводе Радфорд и сарадници (Radford et al. 2023: 28499), стопа грешака на нивоу речи (енг. *word error rate*, скр. *WER*) има јаку негативну корелацију са бројем сати транскрибованог говора за одређени језик ($r^2 = 0.83$).

Такође, када је у питању употреба модела отвореног кода за транскрипцију говора, за њихову употребу потребно је познавање напредних софтверских алата, као и поседовање (или изнајмљивање) значајних техничких ресурса.

Горенаведени фактори су били мотивација за овај рад, у оквиру кога смо представили три различите методе примене верзије модела за аутоматско препознавање говора која је фино подешена за српски језик, чиме се истраживачима и другим корисницима омогућује једноставан, бесплатан и транспарентан приступ овој технологији. Такође, у раду је представљена упоредна анализа перформанси сва три приступа, као и језичка својства саме транскрипције која је произашла из њих.

2. Методологија

2.1. Основна методологија процеса транскрипције

Када је у питању фаза планирања система аутоматске транскрипције говора у нашем истраживању и компоненти које би он обухватао, основну одлуку у овом процесу представљао је одабир самог модела за транскрипцију говора, односно за аутоматско препознавање говора. Као основни модел одабран је *Whisper* модел компаније *OpenAI* (Radford et al. 2023), заснован на кодер-декодер трансформер архитектури, која се, како наводе Сузић и сарадници (Suzić et al. 2024: 175), показала поузданом у литератури и у пракси. С обзиром на мали удео српског говора у оригиналном *Whisper* моделу (Radford et al. 2023: 28517), одабрали смо *whisper-large-v3-sr-cmb*⁴, верзију наведеног модела која је дообучена за српски језик на следећим скуповима података са српским говором: *Google/Fleurs*⁵, *Mozilla/Common Voice*⁶ и *JuzneVesti-SR*⁷ (Сагић 2024: 221).

⁴ <https://huggingface.co/Sagicc/whisper-large-v3-sr-cmb>

⁵ <https://huggingface.co/datasets/google/fleurs>

⁶ https://huggingface.co/datasets/mozilla-foundation/common_voice_9_0

⁷ <http://hdl.handle.net/11356/1679>

Следећа одлука у истраживању тичала се платформе на којој бисмо омогућили читавање и бесплатну употребу овог модела, без потребе за инсталацијом додатног софтвера и без потребе за напредним хардверским ресурсима. Као најбоље решење показала се платформа *Google Colab*⁸, осмишљена као истраживачки пројекат за развој прототипа модела машинског учења, а која пружа временски ограничен приступ окружењу за *Jupyter notebook*⁹, у коме се може вршити интерактивни развој и извршавање програма (Bisong 2019: 59), а чији су хардверски ресурси довољни за покретање различитих модела вештачке интелигенције, укључујући и *whisper-large-v3-sr-cmb*.

Први корак у процесу израде скрипта било је омогућавање читавања аудио-фајлова путем графичког интерфејса, што платформа *Google Colab* подржава помоћу функције *files.upload()* у оквиру своје *Python* библиотеке.

Затим, било је потребно омогућити подршку за различите фајлова и прилагодити их формату који очекује *Whisper* модел, који је трениран на моно аудио-фајловима са фреквенцијом одабирања од 16,000 Hz (Radford et al. 2023: 28494). Ово је постигнуто обрадом улазног аудио-записа помоћу *ffmpeg* библиотеке (преко *pyDub*¹⁰ библиотеке), при чему је, такође, извршена нормализација нивоа звука на -20 dBFS.

Када је у питању сама употреба модела за предикцију (енг. *inference*), *Whisper* модел подржава контекст од 30 секунди, због чега је за сегменте који трају дуже потребно извршити сегментацију (Radford et al. 2023: 28501). У *transformers* библиотеци платформе *HuggingFace*, која подржава позивање *Whisper* модела, постоје два мода¹¹: први, заснован на „исечцима” (енг. *chunks*), у коме се аудио-запис дели на више краћих сегмената са малим преклапањем између сегмената, и други, секвенцијални алгоритам, у коме се аудио-запис дели на сегменте од 30 секунди са „клизајућим прозором” (енг. *sliding window*), који користи транскрипцију из претходног сегмента као додатни контекст, што доприноси већој тачности. У процесу тестирања ова два приступа, други, секвенцијални алгоритам, показао се као тачнији приступ јер је садржао мање грешака на крајевима сегмената у односу на први приступ. Како би крајња транскрипција

⁸ <https://colab.research.google.com/>

⁹ <https://jupyter.org/>

¹⁰ <https://github.com/jiaaro/pydub>

¹¹ <https://huggingface.co/openai/whisper-large-v3>

била што квалитетнија, у нашем раду смо за предикцију помоћу *transformers* библиотеке одабрали секвенцијални приступ.

Приликом анализе транскрипције добијене овим приступом, био је приметан већи број грешака на крајевима реченица, које су типичне за грешке у сегментацији. Из овог разлога, одлучили смо да тестирамо други приступ, у коме бисмо применили напреднији вид сегментације. С обзиром на то да је *Whisper* модел (неправилно) транскрибовао и сегменте говора који су садржали звуке који нису говор (као што је музика), одлучили смо да у новој методологији за транскрипцију применимо моделе за детекцију гласовне активности (енг. *voice activity detection*, скр. *VAD*), који омогућују боље препознавање почетка и краја говора од једноставног приступа заснованог на тишинама у звучном сигналу (Behre et al. 2022; Ding et al. 2020: 433). У овој фази, тестирали смо популарне моделе отвореног кода за детекцију говора као што су *WebRTC VAD*¹², *Silero VAD*¹³, и *Pyannote* (Bredin et al. 2020), при чему је *Pyannote* у нашим тестовима остварио најбоље резултате. Како бисмо омогућили упоредивост са претходним приступом, заснованим на аутоматској сегментацији помоћу *transformers* библиотеке, извршена је иста нормализација нивоа звука као у претходном приступу и задржана иста фреквенција одабирања звука. Такође, како бисмо осигурали компатибилност овог приступа транскрипцији са *Whisper* моделом, применили смо адаптивну сегментацију, у којој су најпре добијени сегменти говора на основу *VAD* модела, а затим је на сваки сегмент који је трајао дуже од 30 секунди примењена функција *split_on_silence* модула *pyDub* уз почетне параметре минималног трајања тишине (*min_silence_len*) од 500 милисекунди, прага тишине (*silence_thresh*) од -40 dBFS, и задржавања сегмента тишине (*keep_silence*) од 200 милисекунди; у случају неуспешне сегментације, вршено је до 5 покушаја аутоматске сегментације уз скраћивање минималног трајања тишине за 100 милисекунди и повећањем прага тишине 2 dBFS. Уколико и након ове процедуре сегментација није произвела сегменте адекватне дужине, сегмент је подељен на најмањи број једнаких делова који су краћи од 30 секунди. Такође, како би се смањила вероватноћа одбацавања речи на самим крајевима сегмента од стране модела, приликом сегментације додато је 100 милисекунди тишине на почетак и крај сваког сегмента. Наведени приступ сегментацији звука довео је до

¹² <https://github.com/wiseman/py-webrtcvad>

¹³ <https://github.com/snakers4/silero-vad>

смањења броја грешака у транскрипцији, што је детаљније описано у другом делу рада.

Упркос бољој сегментацији, добијена транскрипција је и даље садржала значајан број грешака. Како бисмо истражили могућности аутоматске корекције грешака у процесу транскрипције српског говора, креирали смо трећи скрипт, у коме се примењује неколико метода аутоматске детекције и исправљања грешака у транскрипцији. У овом приступу, након што модел генерише транскрипцију, она се најпре прослеђује модулу *CLASSLA-Stanza*, који представља скуп алата за аутоматску језичку анотацију јужнословенских језика (Ljubešić et al. 2024). Помоћу овог модула транскрибовани текст се дели на реченице, а затим се добијене реченице деле на токене помоћу следећег регуларног израза: „ $|w+|/[^\wedge w\backslash s]/|s+|$ ”, који раздваја реченицу на речи, интерпункцију и размаке. Појединачни чланови овог низа се затим могу мењати у исправкама, а *join* функцијом се склапа финална реченица, чувајући оригиналну интерпункцију и размаке.

Један од основних елемената у нашем приступу аутоматској корекцији транскрипције представља листа фреквенција српских речи, коју смо за потребе овог рада креирали на основу два извора:

- морфолошког речника *SrpMD* (Krstev & Vitas 2015), чији су облици речи аутоматски сматрани валидним, и којима су такође на основу *SrpMD* речника придружене информације о врсти речи и леми;
- подкорпуса великих веб-корпуса и великих електронских корпуса српског језика, међу којима издвајамо корпусе Кишобран и Знање (Škorić & Janković 2024), одакле су издвојени облици речи који имају више од 10,000 појављивања у обједињеном корпусу. На основу истих корпуса генерисана је и листа фреквенција биграма у обједињеном корпусу.

Ове листе фреквенција уčitане су у модул *symspellpy*, који је био основа за анализу грешака у транскрипцији и њихово исправљање. Израда наведених листа фреквенција биће детаљније описана у другом раду. У нашој методологији, свака реч из претходног корака упоређивана је са наведеним листама фреквенција.

Како би се избегле непожељне исправке, примењено је неколико врста критеријума који су изузимали реч из процеса аутоматског исправљања грешака. Примери типова речи које су аутоматски изузете су: облици речи који постоје у *SrpMD*

морфолошком речнику, облици речи чија је фреквенција већа од 100, властите именице (речи које је *CLASSLA-Stanza* модул класификовао као *PROPN*), речи које почињу великим словом а не налазе се на почетку реченице, речи које садрже бројеве, познате скраћенице, једно или два велика слова у средини реченице (ради избегавања измене иницијала), речи које садрже латиницу, речи под наводницима и интерпункција и размаци.

За сваку реч која није задовољавала један од ових критеријума, помоћу модула *symspellpy*, а на основу претходно поменути листе фреквенција, генерисани су кандидати за замену речи поређани по растојању уређивања (енг. *edit distance*), а затим по учесталости. Како би се погрешне исправке смањиле на минимум, на наведене кандидате је затим примењено неколико конзервативних филтера, као што је висока сличност између речи, усаглашеност оригиналне речи и предложене замене по врсти речи (према *CLASSLA-Stanza* модулу). Након филтрирања, преостали кандидати ранжирани су према скору који обједињује фреквенцију предложених речи, скор из једноставног биграмског језичког модела и скор којим се фаворизују речи које су већ присутне у тексту.

С обзиром на то да сви претходни механизми исправљања грешака не узимају у обзир контекст, значајан број исправака се састојао од сличних речи које су семантички биле потпуно погрешне. Из овог разлога, за предложене речи чија је фреквенција мања од 500 применили смо методологију исправки која користи модел *Jerteh-355*, који остварује најбоље резултате у моделовању маскираног језика за српски језик (Škogić 2024). У овом кораку, изградња маскиране реченице вршена је заменом циљане речи токеном *<mask>*, а затим је генерисано 15 предлога модела за дату реч (*top_k=15*). Затим, ови предлози су пролазили сличну процедуру филтрирања и поновног рангирања као и речи у претходном приступу, при чему је у овом случају у обзир скор вероватноће који модел пружа, док је максимално растојање уређивања у овом случају било 1^{14} . У финалном кораку, врши се уједначавање предложених речи на основу њихове учесталости у самом тексту. С обзиром на ограничења у простору, у опису методологије дат је преглед само њених најзначајнијих аспеката, док ће више техничких детаља о приступима наведеним у овом раду бити доступно на *GitHub* страници везаној за наше истраживање¹⁵, која ће бити јавно доступна по објављивању рада.

¹⁴ Изузетак је комбинација других фактора који веома позитивно утичу на скор, када је максимално растојање уређивања 2.

¹⁵ <https://github.com/nikolakov/whisper-sr-transcription>

Такође, напомињемо да су у процесу израде методологије, креирању кода и вршењу анализа коришћени модели вештачке интелигенције компанија *OpenAI* и *Google*, при чему смо вршили ручну проверу креираног кода, а предлоге методологије смо валидирали истраживањем релевантне стручне и научне литературе.

3. Квантитативна и квалитативна анализа грешака

Тачност различитих модела транскрипције, описаних у претходним сегментима рада, проверавали смо користећи аудио-записе преузете са платформе *Јутјуб* (енг. *Youtube*). У питању је звучни материјал у трајању од 2 сата и 32 минута, ексцерпиран из видео-записа документарног филма *Магија истока*¹⁶ и кратких предавања из области биологије, преузетих са образовног канала *Изокренута учионица*¹⁷. Целокупан звучни материјал анализиран је у оквиру сваког појединачног модела, те ће анализа грешака бити представљена с освртом на тип транскрипционог модела који је коришћен.

С обзиром на то да модел због „кодер-декодер трансформер архитектуре” емитује токене који могу чинити више од једне фонеме, као и да је токенизатор *Whisper* модела преузет из GPT-2 модела, који је у највећој мери трениран на енглеском језику, упркос подршци за више језика, оваква архитектура модела може довести до различитих грешака и формирања речи које у испитиваном језику не постоје (Radford et al. 2023: 28495; Suzić et al. 2024: 175).

Наш скуп података за валидацију садржао је сегменте са говором дијалекта, као и прекључивање кодова (енг. *code switching*), који су садржали већи број грешака у односу на друге сегменте. Ова чињеница је у сагласности са резултатима претходних истраживања, према којима ове врсте језичке варијабилности доприносе већем броју грешака у транскрипцији (Errattahi et al. 2018: 32).

3.1. Квантитативна анализа грешака

Како бисмо омогућили објективно и конзистентно упоређивање различитих приступа које смо користили, применили смо квантитативну анализу грешака, у којој смо помоћу *Python* скрипта упоређивали финалну верзију транскрипције у сва три приступа. Овај корак обухватао је израчунавање стопе грешке на нивоу речи (енг. *word error rate*, скр. *WER*) упоређивањем ручно

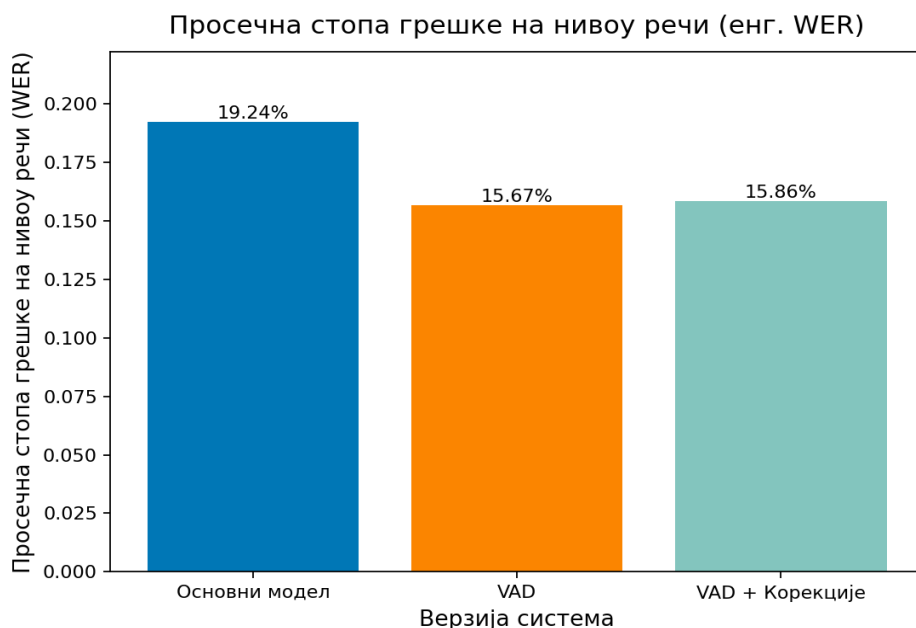
¹⁶ https://www.youtube.com/watch?v=8xAIB_TH3nY

¹⁷ <https://www.youtube.com/user/irinabiologija>

транскрибоване верзије са аутоматски транскрибованим верзијама: $WER = (S + D + I) / N$, односно дељењем збира измењених речи (S), избрисаних речи (D) и уметнутих речи (I) са укупним бројем речи у тексту (N).

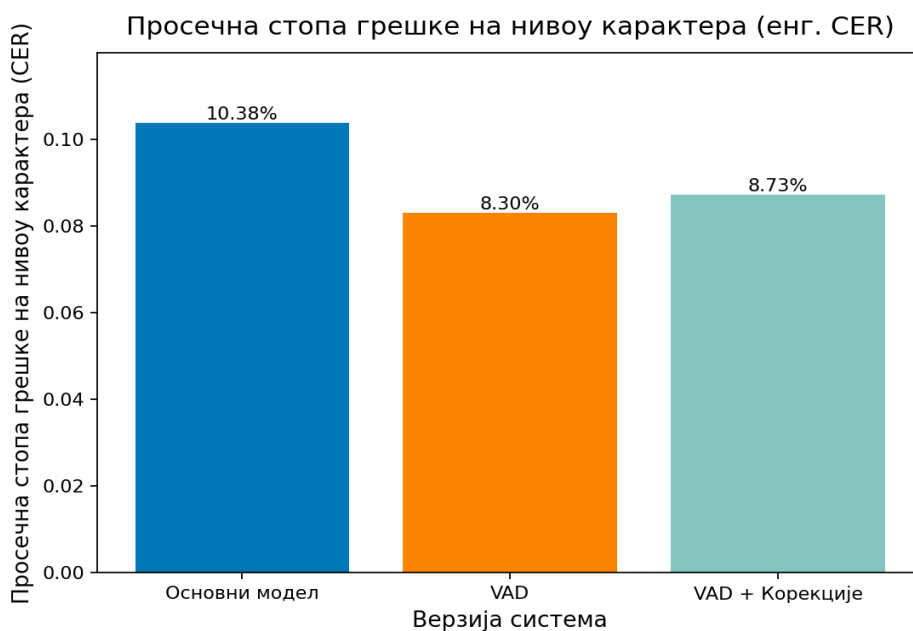
Помоћу *Python* библиотеке *jiwer* вршено је упоређивање грешака само за облик речи (енг. *plain WER*), где је пре упоређивања најпре вршена нормализација уклањањем интерпункције, сажимање размака и конверзија великих у мала слова. Осим грешака на нивоу речи, вршено је и израчунавање стопе грешке на нивоу карактера (енг. *character error rate*, скр. *CER*).

На фигурама 1 и 2 приказана је просечна стопа грешке на нивоу речи (енг. *plain WER*) и на нивоу карактера ¹⁸(енг. *CER*), редом, за сва три приступа.



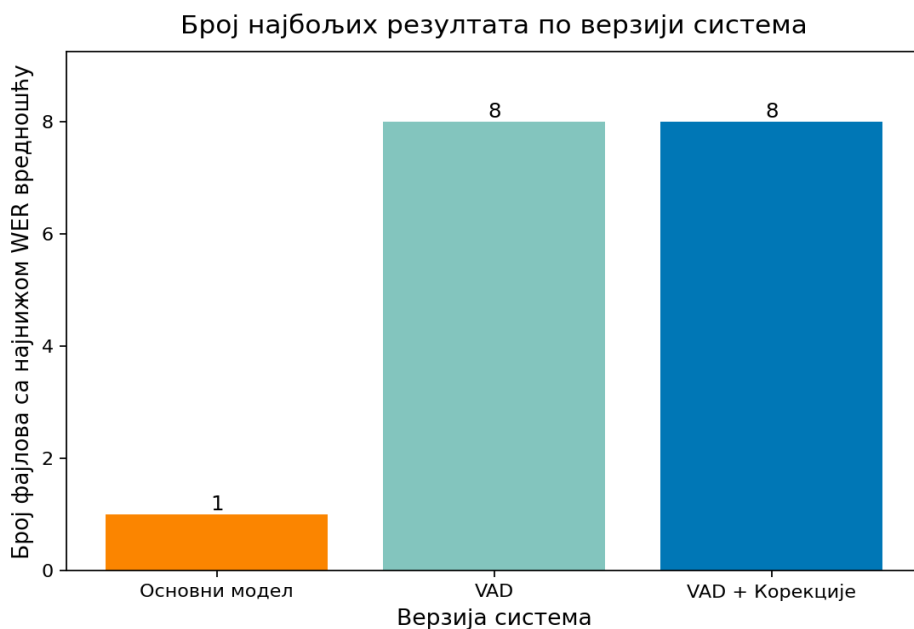
Фигура 1: просечна стопа грешке на нивоу речи (енг. WER) за сва три приступа

¹⁸ <https://github.com/jitsi/jiwer>



Фигура 2: просечна стопа грешке на нивоу речи (енг. WER) за сва три приступа

Као што се може видети на фигурама 1 и 2, наша два модификована приступа остварила су значајно боље резултате у односу на оригиналну сегментацију у подразумеваној процедури за транскрипцију *transformers* библиотеке. Међутим, најмањи број грешака (како на нивоу речи, тако и на нивоу карактера) забележен је са *VAD* моделом без корекција транскрипције (15.67% на нивоу речи и 8.3% на нивоу карактера), док је приступ са корекцијама транскрипције имао нешто већу просечну стопу грешака (15.86% на нивоу речи и 8.73% на нивоу карактера). Са друге стране, као што се може видети на фигури 3, од 17 фајлова у валидационом скупу података, оба приступа са нашим модификацијама имала су исти број фајлова у коме су остварила најбољи резултат од сва три тестирана (при чему напомињемо да дужина фајлова није узета у обзир). Ови подаци указују на то да примењена методологија исправљања грешака у сегментацији у неким типовима текста доводи до бољих резултата, али да и даље постоји значајан број нежељених исправки које у одређеним случајевима доводе до већег укупног броја грешака.



Фигура 3: Број фајлова у којима је одређена верзија имала најмању стопу грешке у односу на остале верзије

С обзиром на то да WER не осветљава појединачне језичке аспекте који су специфични за различите приступе аутоматској транскрипцији (и аутоматским исправкама које се над њом врше), поред квантитативне анализе, извршили смо и квалитативну анализу транскрибованог текста добијеног у сва три приступа.

3.2. Квалитативна анализа грешака

3.2.1. Грешке на нивоу речи или фразе

Највећи недостатак претварања говора у писани текст коришћењем ових модела представља изостанак одређених речи. Иако би коришћење ових модела знатно умањило време које лингвисти користе за транскрипцију, изостављање делова текста онемогућило би коришћење ових модела за области у којима је важна свака изговорена реч, попут прагматике, фонетике, дијалектологије итд.

Речи и фразе које модел углавном прескаче јесу такозвани *језички испуњивачи*, речи које попуњавају паузе и празнине у говору, у тренуцима када говорник покушава да осмисли свој говор, не може да се сети шта је желео да каже, жели да исправи нешто у свом говору

или га допуни, као што су *неуобличена вокализација, поштапалице, понављање, надопуњавање* итд. (Кликовац, Чудомировић 2016: 7). Дакле, у потпуности су изостављене поштапалице, попут *значи, мислим, овај, па, некако, овако*, које, иако понекад означене као *језички шунд* (Фекете 1997), указују на важну комуникативну функцију и сигнализирају да говорник размишља о планирању свог исказа, па самим тим нису ни комуникативно сувишне (Чудомировић 2020: 52). Уочљиво је и изостављање дискурских маркера, као што су: *такође, у ствари, међутим, заправо, наравно, управо, односно, рецимо, иначе, дакле* итд. Такође, исправљање говорника модел транскрибује тако што узима само један облик: *па на неким неком > на неком; у јако хладним у јако сувим пределима > у јако сувим пределима; највећи највећа маса > највећа маса*.

Поред тога, модел тежи економичности исказа, те изоставља, сажима и скраћује поједине речи и конструкције: то је било **више** практиковање > то је било практиковање; између Тимока, **исто тако** и Дунава > између Тимока, **али** и Дунава; нећу **чак то** ни на телевизији > нећу ни на телевизији итд. У таквим ситуацијама најчешће се изостављају неодређене и показне заменице (неки, некако, овај, онај), показне речце којима се уводе појмови (ево, ето, ено) и већ поменути испуњивачи. Имајући у виду особине скупова података на којима се модели за транскрипцију говора најчешће тренирају, а у којима се тежи језичкој економији, са једне стране, а да процес сегментације може довести до грешака у транскрипцији на крајевима појединачних звучних исечака, са друге стране, није јасно да ли су одређене грешке узроковане самим својствима модела или неправилном сегментацијом (некадашњу ботаничку башту > неботаничку башту; ако ћемо још да идемо > хућемо да идемо).

3.2.2. Грешке на нивоу гласа или групе гласова

Понекад због недовољно развојеног говора, затвореног или отвореног изговора вокала, долази до супституције, редукције или пак додавања гласа који није изговорен. Разлоге за то можемо потражити у поставкама *Whisper* модела, будући да су токени којима он располаже, а који могу садржати једну или више фонема, преузети из одређеног *ChatGPT* модела, те тако настају речи које у српском језику не постоје (Suzić et al. 2024: 175). Стога, како би се избегле грешке које подразумевају речи које не постоје у српском језику, у трећи модел транскрипције имплементиран је *фреквенцијски речник*, чија је методологија примене описана у претходном делу рада.

Примери који указују на то да је имплементирање фреквенцијског речника допринело тачности транскрипције јесу оне речи које су у прва два типа транскрипције, оригиналном моделу и моделу са ВАД сегментацијом, транскрибоване са одређеним фонетским грешкама, док у трећем моделу таквих грешака нема: *екстремитети* > *екстримитети* (1), *екстримитети* (2), *екстремитети* (3); *кичмењацима* > *кичмејацима* (1), *кичмејацима* (2), *кичмењацима* (3). Међутим, иако је имплементација модела повећала тачност транскрипције, постоје и речи које су тачне у оригиналном типу транскрипције, а погрешне у другом и трећем, попут *оповргнути* > *оповргнути* (1), *поврнути* (2), *оповркнути* (3).

Будући да правилности у појављивању грешака у гласовним сегментима нема, тешко је одредити који тип грешака је учесталији, те да ли модел више греша када су у питању вокали или консонанти. Прво, када су у питању вокали, грешке у транскрипцији су бројне, што понекад можемо приписати неадекватном изговору говорника. Честе су супституције (*архетипске* > *архитипске*, *круне* > *круну*, *хладноће* > *хладноћа*, *попречном* > *попричном*) и редукције вокала (*смисао* > *смиса*), а понекад долази и до уметања или удвајања вокала (*почнимо* > *починимо*, *усред* > *уусред*). Грешака има и међу консонантима, посебно уколико су у питању речи у којима је извршена нека гласовна алтернација: *ишчупају* > *исчупају* (1), *ишчупају* (2, 3); *лисну дршку* > *листну дршку* (1), *листну дршку* (2), *листу дршку* (3). Нарочито се истичу примери супституције одређеног гласа његовим звучним односно безвучним парњаком: *искустава* > *изгустава*, *хладноће* > *хлатноће*, *уздужном* > *устужном*, *праином* > *брајантом*. Додатне примере грешака наводимо у табели:

Табела 4: Примери грешака с освртом на тип транскрипционог модела

Оригинална верзија	Whisper (1)	VAD (2)	VAD + Jerteh (3)
котлина	кутлина	кутлина	кутлина
смањују	смањају	смањују	смањују
изблиза	изблизан	изблизак	изблиза
клови	кловен/клоун	кловен/клоун	кловен/клоун
сведоче	седоче	следоче	следоче
аутохтона	аутолтона	аутохтона	аутоктона
багрем	багрем	багрѐн	багрѐн
кестен	кестен	кестер	кеспен
поменули	упоменули	упоменули	поменули
отров	отворов	отворов	отвор пров
успут	успутом	успутом	успутном
кисеоник	кисони	кисони	кисони
спољашње	спољашће	спољешће	спољашће
причвршћен	причајћен	причај	причај

Истоветан говорни сегмент понекад се транскрибује исправно, а некада није могуће одгонетнути његово право значење. Пример *што зглавкара чини зглавкар*ом понавља се два пута у току аудио-записа. Први пут резултат транскрипције јесте: *спод изглавкара, чини изглавкар*ом (1), *спод злавка рачини и злавка ром* (2), *спод зглавка рачини и зглавка рома* (3), док је сличан сегмент, када је други пут поменут, исписан тачно: *шта то зглавкара чини зглавкар*ом (1, 2, 3). Такође, реч *нерватура* први пут је транскрибована као: *на ратвору* (1, 2), *на раствору* (3), док други пут имамо тачну реч – *нерватуру* (1, 2, 3).

Уколико је одређена реч кандидат за исправљање, а само се у једном гласу разликује од неке речи која постоји у фреквенцијском речнику, модел ће преузети ту реч, иако је семантички потпуно различита, што представља ману овог приступа: *врхова* > *вркова* (1), *вркована* (2), *врбована* (3); *науколике* > *околики* (1), *околики* (2), *околини* (3); *сгазите* > *схазите* (1), *сгазите* (2), *спазите* (3). Како би се избегао овакав тип грешака, било би пожелено, у наредним имплементацијама модела, применити напреднији систем корекције који би узимао у обзир контекст. Са друге стране, предност модела увиђа се када модел уноси реч која је у оригиналу непотпуна: *специ...* за *кретање* > *специјално* за *кретање*.

Неадекватна је и употреба великог слова, будући да модел понекад адекватно препознаје примере у којима је велико слово неопходно (*Дунав*, *Србија*, *Венера*), понегде погрешно наводи мало (*Бечу* > *бечу*, *Рајац* > *рајац*, *Перуна* > *перуна*), а понегде велико слово (*пирамиде* > *Пирамиде*, *српски и словенски* > *Српски и Словенски*, *чудо* > *Чудо*) итд.

Када је у питању транскрипција *антропонима* и *топонима*, модел у све три верзије подједнако греши, с обзиром на то да су властите именице изостављене из исправки. Примери попут *Дунав*, *Србија*, *Балкан*, *Венера*, *Дионис*, *Хиландар*, *Ирина* и сл. исправно су транскрибовани, уз поштовање правила о великом слову, док типичне грешке увиђамо код презимена и појединих топонима: *Дамњановић* > *Дамљановић*, *Демљановић*, *Бемљановић*; *Минх* > *Мин*, *Мих*; *планине Ртањ* > *Плање Нертањ*; *Зајечару* > *Зајачару*, а понекад је реч која означава име потпуно изостављена: *хотел Рамонда* > *хотел*.

Интересантна је, такође, транскрипција изговора на енглеском језику, будући да су речи махом транскрибоване тако да подражавају српски изговор енглеских речи, попут *vein* > *вејн*; *phylum* > *фајлом*; *flower* > *флауер*; *midrib* > *мид риб* (1, 2), *миди рибе* (3), *petiole* > *петијал* (1), *петијол* (2, 3); *ovary* > *овери* (1), *оувери* (2), *овери* (3) и др., уз изузетак појединих речи које су преведене или постоје у сличном облику у српском језику (*water* > *вода*, *transports* > *транспорт*, *minerals* > *минерал*).

3.2.3. Грешке на нивоу сегментације текста

Када су у питању адекватна сегментација текста и поштовање интерпункцијских правила српског језика, сви модели показали су добре резултате. Најфреквентније грешке су: недостатак размака између појединих реченица, иако су оне одвојене тачком на крају прве и великим словом на почетку друге реченице, употреба малог слова на почетку реченице и неадекватна употреба зареза, премда су синтаксичке функције које се у српском језику одвајају зарезима углавном адекватно сегментирани: *причамо о листу, биљном органу*.

Поред поменутих одступања, речи су понекад подељене на два дела или су две речи спојене у једну: *сунђери* > *су ђери*, *суњери*, *су њуђари*; *маишту свих путника* > *маиштус их путника*; *јутарњи сунчеви зраци* > *јутарњисунчеви зраци*; *са Ртња* > *свртња* итд. Наведене грешке могу бити узроковане погрешном сегментацијом од стране *VAD* модела, једноставнијом логиком за сегментацију, која је активирана уколико је одређени сегмент био дужи од 30 секунди и након употребе *VAD* модела, или особинама самог *Whisper* модела.

4. Закључак

У овом раду описали смо три различита приступа транскрипцији српског језика уз помоћ Гугл Колаб платформе и *whisper-large-v3-sr-cmb* модела, при чему смо постигли значајно смањење у просечној стопи грешке у односу на подразумевани приступ са *transformers* библиотеком. Сматрамо да приступ сегментацији са *VAD* моделом примењен у овом истраживању представља искорак у транскрипцији говора на српском језику када су у питању решења отвореног кода. Са друге стране, анализе везане за аутоматске исправке транскрипције указују на то да су за поуздан систем аутоматског исправљања грешака потребне методе које би узимале контекст у обзир, затим пречишћене и проширене листе фреквенција, као и експерименти са различитим комбинацијама параметара алата описаних у овом раду. Такође, бројне грешке у иницијалној транскрипцији указују на то да је неопходно додатно обучавање модела за препознавање говора за српски језик. Упркос овим изазовима, сматрамо да наведени приступи транскрипцији могу бити корисни истраживачима, али и другим корисницима за потребе транскрипције српског језика.

Литература

- [1] Behre, P., Parihar, N., Tan, S. S., Shah, A., Sharma, E., Liu, G., Chang, S., Khalil, H., Basoglu, C., & Pathak, S. D. (2022). Smart Speech Segmentation using Acousto-Linguistic Features with look-ahead. *ArXiv, abs/2210.14446*.
- [2] Bisong, E. (2019). Google Colaboratory. In *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners* (pp. 59–64). Apress. https://doi.org/10.1007/978-1-4842-4470-8_7
- [3] Bredin, H., Yin, R., Coria, J. M., Gelly, G., Korshunov, P., Lavechin, M., Fustes, D., Titeux, H., Bouaziz, W., & Gill, M.-P. (2020). Pyannote.Audio: Neural Building Blocks for Speaker Diarization. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7124–7128. <https://doi.org/10.1109/ICASSP40776.2020.9052974>
- [4] Ding, S., Wang, Q., Chang, S.-Y., Wan, L., & Moreno, I. L. (2020). Personal VAD: Speaker-Conditioned Voice Activity Detection. *The Speaker and Language Recognition Workshop (Odyssey 2020)*, 433–439. <https://doi.org/10.21437/Odyssey.2020-62>
- [5] Errattahi, R., Hannani, A. E., & Ouahmane, H. (2018). Automatic Speech Recognition Errors Detection and Correction: A Review. *Procedia Computer Science*, 128, 32–37. <https://doi.org/10.1016/j.procs.2018.03.005>

- [6] <https://api.semanticscholar.org/CorpusID:253116616>
- [7] Jurafsky, D., & Martin, J. H. (2025). Automatic Speech Recognition. In *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models* (3rd ed.). <https://web.stanford.edu/~jurafsky/slp3/15.pdf>
- [8] Krstev, C., & Vitas, D. (2015). *SrpMD — Serbian Morphological Dictionaries*. <https://doi.org/10.57771/j0ge-8e29>
- [9] Ljubešić, N., Rupnik, P., & Koržinek, D. (2025). The ParlaSpeech Collection of Automatically Generated Speech and Text Datasets from Parliamentary Proceedings. In A. Karpov & V. Delić (Eds.), *Speech and Computer* (pp. 137–150). Springer Nature Switzerland.
- [10] Ljubešić, N., Terčon, L., & Dobrovoljc, K. (2024). *CLASSLA-Stanza: The Next Step for Linguistic Processing of South Slavic Languages*. Institute of Contemporary History. <https://doi.org/10.5281/zenodo.13936406>
- [11] McFee, B., Raffel, C., Liang, D., Ellis, D., Mcvicar, M., Battenberg, E., & Nieto, O. (2015). *librosa: Audio and Music Signal Analysis in Python*, 18–24. <https://doi.org/10.25080/Majora-7b98e3ed-003>
- [12] Pratap, V., Xu, Q., Sriram, A., Synnaeve, G., & Collobert, R. (2020). MLS: a large-scale multilingual dataset for speech research. *Interspeech 2020*. <https://doi.org/10.21437/interspeech.2020-2826>
- [13] Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In A. Celikyilmaz & T.-H. Wen (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 101–108). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-demos.14>
- [14] Radford, A., Kim, J. W., Xu, T., Brockman, G., Mcleavey, C., & Sutskever, I. (2023). Robust Speech Recognition via Large-Scale Weak Supervision. *Proceedings of the 40th International Conference on Machine Learning*, 28492–28518. <https://proceedings.mlr.press/v202/radford23a.html>
- [15] Suzić, S., Popović, B., Pakoci, E., Pekar, D., Delić, T., & Delić, V. (2024). *Automatic Speech Recognition in the Serbian Language – A Comparison of Whisper-Based System and Conventional System with Language Model*, 174–177.
- [16] Škorić, M., & Janković, N. (2024). New Textual Corpora for Serbian Language Modeling. *Infotheca*, 24(1), 71–96. <https://arxiv.org/abs/2405.09250>
- [17] Škorić, M. (2024). New Language Models for Serbian. *Infoteka*, 24(1), 7–28. <https://arxiv.org/abs/2402.14379>

*

- [18] Кликовац, Д., & Чудомировић, Ј. (2016). Скрипта из Лингвистике текста и прагматике за студенте српског језика и књижевности.
- [19] Сагић, А. (2024). Употреба модела вештачке интелигенције у процесу дигитализације аудио и видео-грађе. Гласник Народне Библиотеке Србије, 23(26), 217–229.
- [20] Фекете, Е. (1997). Поштапалице – елементи језичког шунда. Језик данас, 4, 5–8.
- [21] Чудомировић, Ј. (2020). Зашто су „поштапалице” потребне. Књижевност и језик, 67(1), 41–55.

Usage of the *Whisper Large v3 Sr* model for the transcription of serbian spoken language in *Python* programming language on the *Google Colab* platform

Nikola Janković, Jovana Ivaniš

Summary

This paper presents a *Python* script on the *Google Colab* platform, which uses a fine-tuned model for the transcription of speech in Serbian, *Whisper Large v3 Sr* (<https://huggingface.co/Sagicc/whisper-large-v3-sr-cmb>), which enables free, high-quality and simple transcription of Serbian speech into text. The motivation for creating this script came from the lack of available tools which would allow researchers to use this model in a straightforward way, without the need for advanced technical knowledge and significant computing resources. This script provides a simple method for uploading audio files, transcribing them using the *Whisper Large V3 Sr model*, and downloading the transcription of the files in a textual format. Firstly, the paper briefly describes the aforementioned model used by the script, followed by a detailed description of how the script functions, along with a user manual. The paper will also present problems which we needed to overcome in order to successfully implement this approach, such as the need for automatic audio file segmentation, the determination of the optimal segmentation parameters, and the implementation of support for different audio file formats. Furthermore, several approaches related to the reduction in the number of errors in the transcription will be presented. We believe that this tool can be of significant importance to researchers,

considering that it speeds up the processing of audio data, enables users to process vast amounts of audio material in a short period of time, and provides a consistent and repeatable transcription method, which is very significant for the scientific methodology and the repeatability of language-related research. We also believe that the tool can be useful to other users, considering the fact that it enables creating subtitles for video content, converting audio notes into text, creating automatically generated captions for the deaf and hard of hearing, as well as creating textual archives of audio content.

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

Od rječnika kao korpusa do hiponimije i meronimije¹

Naučni rad

DOI: 10.18485/judig.2025.1.ch28

Maja Matijević²,  0000-0003-3990-0826

Apstrakt

Rad se bavi hijerarhijskim leksičko-semantičkim odnosima, hiponimijom i meronimijom, te poluautomatskim prepoznavanjem tih odnosa u rječničkim definicijama. Hiponimija se pritom definira kao odnos između leksičke jedinice koja označuje vrstu i leksičke jedinice koja označuje predstavnika vrste (npr. hiperonim *životinja* > hiponim *pas*), a meronimija kao odnos između leksičke jedinice koja označuje cjelinu i leksičke jedinice koja označuje dio te cjeline (npr. holonim *tijelo* > meronim *ruka*). Definira se teorijski okvir te se prikazuje mogućnost automatiziranoga prepoznavanja u rječničkim definicijama. Metodološko polazište čine leksičko-sintaktički obrasci, tj. tipične strukture s određenim sintaktičkim pravilnostima i leksičkim jedinicama s pomoću kojih se mogu prepoznati leksičko-semantički odnosi. Analizom rječničkih definicija dobiveno je više obrazaca te su istaknuti padežni i prijedložni izrazi, važni za izricanje odnosa dijela i cjeline. U radu se ističe da su rječnici zbog ujednačene strukture bogat i pouzdan izvor za prepoznavanje hijerarhijskih odnosa, što otvara mogućnost za daljnja jezična istraživanja i analize.

Ključne riječi: leksičko-semantički odnos, hiponimija, meronimija, leksičko-sintaktički obrazac, rječnik

¹ Ovaj rad napisan je na istraživačkome projektu *Hrvatski mrežni rječnik (Mrežnik): 2. faza*, koji financira Europska unija – NextGenerationEU. Za iznesene stavove i mišljenja odgovorna je samo autorica te ti stavovi ne odražavaju službena stajališta Europske unije ili Europske komisije. Ni Europska unija ni Europska komisija ne mogu se smatrati odgovornima za njih.

² Institut za hrvatski jezik, mmatijevic@ihjj.hr

1. Uvod

Hiponimija i meronimija hijerarhijski su leksičko-semantički odnosi, tj. odnosi koji ustrojavaju leksičke hijerarhije. Hiponimija je pritom odnos između leksičke jedinice koja označuje vrstu općenito i predstavnika te vrste (npr. *životinja* > *pas*), a meronimija odnos između leksičke jedinice koja označuje cjelinu i leksičke jedinice koja označuje dio te cjeline (npr. *tijelo* > *ruka*).

Glavna je značajka obaju odnosa višerazinski ustroj, iz čega proizlazi to da leksičke jedinice u tim odnosima nisu zrcalne, simetrične (kao što su primjerice dva antonima jedan drugomu antonimi, npr. *crn* – *bijel*). Točnije, leksičke jedinice koje su na višoj razini u hiponimijskoj hijerarhiji nazivaju se hiperonimi (npr. *životinja*), a jedinice na nižoj razini hiponimi (*pas*, *mačka*). Dva hiponima istoga hiperonima nazivaju se kohiponimi (*pas* – *mačka*). Leksička jedinica koja u meronimijskome odnosu označuje cjelinu naziva se holonim (npr. *tijelo*), a jedinica koja označuje dio meronim (*ruka*, *noga*). Dva meronima istoga holonima nazivaju se komeronimi (*ruka* – *noga*).

Hijerarhijske odnose tematiziraju još grčki filozofi, zatim srednjovjekovni mislioci, a pitanja poput djeljivosti ostaju važna i u daljnjim epohama filozofske misli (usp. Materynska 2019: 256). Odnosi su važni i u antropološkim radovima, računalnoj znanosti (primjerice u sustavima upita i odgovora), no dugo nisu bili opisivani u jezikoslovlju. Hiponimija i meronimija počinju se sustavnije proučavati 80-ih i 90-ih godina, i to najviše u psiholingvističkim radovima koji se bave značajkama paradigmi, npr. meronimijskih (Winston i dr. 1987, Iris i dr. 1988, Gerstl i Pribbenow 1996 i dr., usp. Materynska 2019: 256).

Naglasak na proučavanje leksičko-semantičkih odnosa u jezikoslovlju se stavlja od samih početaka strukturalizma, 30-ih godina 20. stoljeća (Geeraerts 1999: 126, Geeraerts 2010: 49–51), no sustavno se opisuju krajem strukturalizma, 60-ih godina (npr. Lyons 1963, Lyons 1977). Leksičko-semantički odnosi računalno su najdetaljnije i najsustavnije povezani u semantičkim mrežama odnosno računalnim tezaurusima, među kojima se osobito ističe *WordNet* (Miller 1990, 1995, Fellbaum 1998). Skupovi sinonima koji predstavljaju jedno značenje (sinseti) povezani su u različite odnose: antonimija, troponimija (među glagolima), uzročnost, logička nužnost te, vrlo zastupljeni i oprimjereni odnosi, hiponimija i meronimija. Pritom se odnosi dijele na podvrste, pa postoji meronimija u kojoj je dio (1) komponenta cjeline, (2) član cjeline ili (3) materijal cjeline (tvar od koje je cjelina načinjena) (Miller 1990:

256). Semantičke mreže razvijaju se i za južnoslavenske jezike, npr. *CroWn* za hrvatski, *SrpWn* za srpski, *BulNet* za bugarski, *SloWNet* za slovenski.

Među leksičko-semantičkim odnosima općenito više se proučavaju sinonimija i antonimija, o čemu svjedoči i broj objavljenih monografija. Tako je primjerice u hrvatskome jezikoslovlju objavljena monografija o sinonimiji (Petrović 2005) te rječnik sinonima (Šarić i Wittschen 2008), cjelovita monografija napisana je i o antonimiji (Šarić 2007), a obuhvatni radovi o hiponimiji i meronimiji u hrvatskome jeziku, a ni u južnoslavenskim jezicima općenito, nisu objavljeni.

Ovaj rad definira dva važna leksičko-semantička odnosa i daje uvid u mogućnosti poluautomatskoga izlučivanja primjera iz rječničkoga korpusa koji zatim mogu poslužiti za dodatne analize. Rad se oslanja na postavke distribucijske semantike, tj. polazi od hipoteze da se značenje riječi može prepoznati prema kontekstu. Točnije, glavno su metodološko polazište leksičko-sintaktički obrasci (tako nazvani u Hearst 1992) kao tipične strukture prema kojima se mogu prepoznati parovi hiperonima i hiponima te holonima i meronima. U sljedećemu poglavlju detaljnije će se definirati leksičko-sintaktički obrasci te će se dati pregled proučavanja obrazaca. Treće poglavlje donosi osnovne informacije o korpusu koji se analizira, a u četvrtome donose se obrasci za hiponimiju i meronimiju.

2. Leksičko-sintaktički obrasci

Najutjecajniji predstavnici leksičke semantike u svojim kapitalnim radovima ističu da ostvareni obrasci daju naznake određenoga odnosa (Lyons 1977: 311) odnosno da postoje *dijagnostički okviri* ili *dijagnostičke formule* (Cruse 1986: 12). Cruse navodi da dijagnostički okviri služe za prepoznavanje značenja na osnovi distribucije i oblika te ih uspoređuje s lakmusovim papirom – čovjek ne može odrediti alkalnost tekućine, ali može procijeniti je li lakmusov papir plav ili crven i na temelju toga donijeti zaključak. U jeziku tako postoje dijagnostički okviri koji upućuju na što, pojednostavnjuju intuiciju i lakše dovode do zaključaka (Cruse 1986: 10–11). Upravo su hijerarhijski odnosi poslužili kao prvi primjer odnosa u kojima postoje dijagnostički okviri, odnosno Cruse (1986: 13, također i 2000: 153) najprije navodi okvir *X i ostali Y* (engl. *X and other Y*) za hiponimiju i *X je dio Y-a* (engl. *X is a part of Y*) za meronimiju.

Sljedećih godina dolazi do zamaha u proučavanju okvira, pa valja istaknuti da Justeson i Katz (1991) istražuju okoline u kojima se pojavljuju antonimi, a postoje i brojna druga istraživanja sintaktičkih okruženja u kojima se mogu pojaviti antonimi (npr. Jones i Murphy 2005, Jones 2006, Jones 2010, v. i Čulig Suknaić 2020: 57–58).

Ipak, u ovome kontekstu najvažniji rad objavljuje Hearst (1992), koja razvija metodu prepoznavanja hiponima s pomoću čestih *leksičko-sintaktičkih obrazaca*. Upućuje na to da se preko sintagmatskoga ostvaraja može doprijeti do paradigmatiskih značenjskih odnosa, a time i do razumijevanja rječničkoga ustroja. Točnije, apstrahira rečenice na leksičko-sintaktičke obrasce – primjerice iz rečenice *The bow lute, such as Bambara ndang, is plucked and has (...)*. (hrv. *Lutnja za gudalo, kao što je bambara ndang, trzalačka je i ima (...)*.) izlučuje obrazac NP_0 *such as* $\{NP_1, NP_2 \dots (and/or)\} NP_n$ (Hearst 1992: 539). Dakle, iz rečenice i obrasca može se zaključiti da je *bambara ndang* hiponim gudalu čak i ako se ne zna što je *bambara ndang*. Leksičko-sintaktički obrazac pretpostavlja da na određeno mjesto dolazi hiponim i na određeno mjesto hiperonim, tj. da određene leksičke jedinice otvaraju mjesto hiponimima, hiperonimima ili kojoj drugoj jedinici, ovisno o strukturi obrasca. Sintaktički gledano, leksička jedinica može doći u točnome padežu, glagol može imati različitu rekciju i sl., što je osobito važno u flektivnim jezicima kao što je hrvatski i drugi južnoslavenski jezici. S obzirom na navedeno i na ustaljenost naziva *leksičko-sintaktički obrazac* – njime se koriste i brojni drugi kasniji autori, npr. Ando i dr. (2004), Snow i dr. (2005), Girju i dr. (2006), Barbu Mititelu (2011), Yıldız i dr. (2016), Sahin (2017), Farrugia (2019) – u ovome radu prednost se daje tomu nazivu.

Leksičko-sintaktički obrasci za hiponimiju (kao i za meronimiju) mogu se izlučivati iz različitih vrsta tekstova. Hearst (1992) se u svojem pionirskom radu služi Grolierovom američkom enciklopedijom, a nadalje primjerice Rydin (2002) analizira novinske vijesti na švedskome, Ando i dr. (2004) analiziraju vijesti na japanskome. Hiponimno-hiperonimni parovi analiziraju se i u kineskome (Liu i dr. 2006), nizozemskome (Sang i Hofmann 2007), rumunjskome (Barbu Mititelu 2011), litavskome (Bielinskienė i dr. 2012), postoje brojni radovi za turski (npr. Yıldırım i Yıldız 2012a, 2012b, Sahin i dr. 2016) i španjolski (Gil-Berrozpe 2017 i Gil-Berrozpe i dr. 2017). Radovi postoje i za slavenske jezike – npr. Przepiórkowski i dr. (2007) za bugarski, češki i poljski, Baisa i Suchomel (2015) za češki, Vintar i Gorjanc (2003) za slovenski. Stanković i dr. (2021) iz velikoga broja definicija izlučuju pravilnosti (vrste definicija te leksičke i sintaktičke značajke) te (polu)automatskom analizom dolaze i do leksičko-semantičkih odnosa među kojima su hiponimija i meronimija. Vrijedan je doprinos u tome radu klasifikacija različitih rječničkih definicija (Stanković i dr. 2021: 698). Kad je riječ o meronimiji, obrasce spominju već Winston i dr. (1987), Iris i dr. (1988), Berland i Charniak (1999), Girju i dr. (2006). Obrasci za prepoznavanje meronimije također se istražuju u brojnim jezicima, također mnogo za turski jezik (v. npr. Yıldız i

dr. 2016). Obrasci, dakako, ovise o jeziku i mogu biti vrlo raznoliki, o čemu svjedoče primjeri iz prikazani na Slici 1.

1	NP _X PP _Y — PP _Y starts with <i>of</i>	173	door of his car the executive of the new government
2	NP _X PP _Y — PP _Y starts with <i>in</i>	61	people in the world
3	NP _X PP _Y — NP _X ends with <i>branch</i> — PP _Y begins with <i>of</i>	14	They organized the executive branch of government
4	NP _X PP _Y — PP _Y starts with <i>from</i>	9	oxygen from air people from all over the world
5	NP _X PP _Y — PP _Y starts with <i>throughout</i>	6	people throughout the world
6	NP _X PP _Y — PP _Y starts with <i>at</i>	5	window at the rear of the building
7	NP _X PP _Y — PP _Y starts with <i>on</i>	16	five fingers on one hand
8	NP _X PP _Y — PP _Y starts with <i>to</i>	3	the door to the house
9	NP _X PP _Y — PP _Y starts with <i>around</i>	2	people around the world
10	NP _X PP _Y — PP _Y starts with <i>all over</i>	1	people all over the world

Slika 1. Neki leksičko-sintaktički obrasci za meronimiju u engleskome i primjeri za svaki obrazac (Girju i dr. 2006: 122)

Za izlučivanje leksičko-sintaktičkih obrazaca dobru podlogu čine rječnici, koji s jasnom strukturom definicija nude mogućnost za (polu)automatsko izlučivanje leksičkih jedinica koje su u hiponimijskome ili meronimijskome odnosu. Za to su osobito korisni mrežni rječnici s dostupnim definicijama, a jedan je od njih za hrvatski jezik *Hrvatski mrežni rječnik – Mrežnik*.

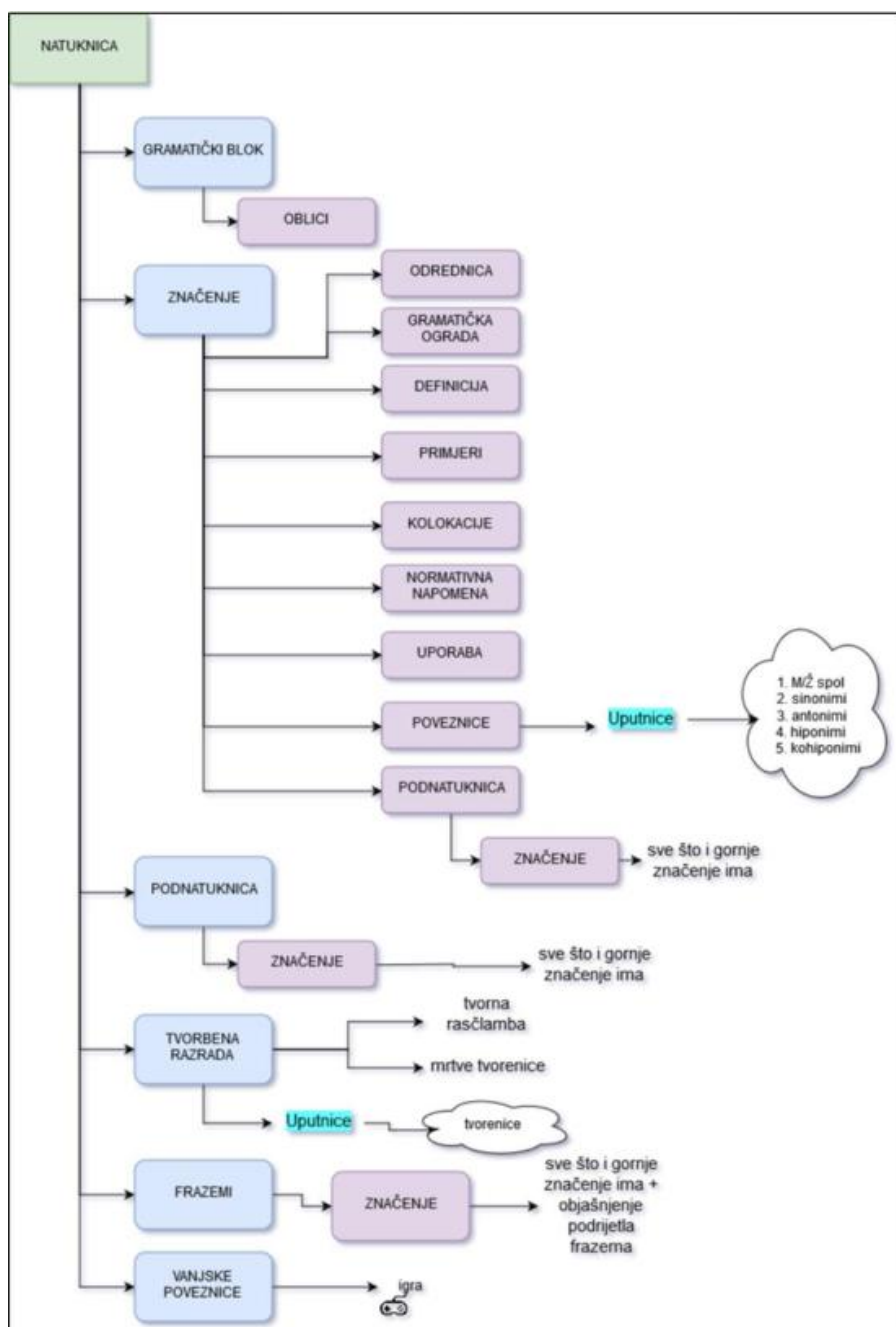
3. Korpus za istraživanje

Rječnička definicija kao jedan od glavnih dijelova rječničkoga članka sadržava podatke o leksičko-semantičkim odnosima, a Zgusta (1971: 253–254) navodi da su u njoj važni upravo hiperonim i potom značenjska razlika, ono što su u logičkoj definiciji *genus proximum* i *differentia specifica*. Osim u rječnicima važnost definicije ističe se u jeziku općenito – definicija (definiranje) osnovna je jezična aktivnost jer njome jezik opisuje sam sebe (Barnbrook 2002). Još se 80-ih godina prošloga stoljeća autori služe rječnicima kako bi došli do leksičko-semantičkih odnosa (npr. Amsler 1980, Markowitz i dr. 1986, Ahlswede i Evens 1988, Iris i dr. 1988), a u novije doba navode da je rječnik dobar korpus za takva istraživanja jer se (1) sastoji od definicija koje pružaju mnogo informacija,

(2) rječničke definicije pišu na isti način, tj. imaju istu strukturu te zbog toga jasne i česte sintaktičke obrasce (Šerbetći i dr. 2011: 13). Stanković i dr. (2021: 698), osim toga, ističu da modeli rječničkih definicija i njihove lokalne gramatike omogućuju povezivanje drugih leksičkih resursa (u njihovu slučaju za srpski jezik) te izdvajanje definicija iz strukovnih korpusa.

Radi lakšega pristupa podacima kao glavni korpus za ovo istraživanje poslužio je prvi mrežni rječnik hrvatskoga jezika: *Hrvatski mrežni rječnik – Mrežnik*.³ U vrijeme pisanja ovoga rada (u srpnju 2025. god.) *Mrežnik* sadržava više od 10 000 natuknica, i to svih vrsta riječi. Natuknica se u osnovnome modulu sastoji od gramatičkoga bloka, značenjskoga bloka, podnatuknica, tvorbenoga bloka, frazema, etimološke obrade i poveznica (više o *Mrežniku* v. u Hudeček i dr. 2024). Cijela struktura natuknice prikazana je na Slici 2.

³ *Hrvatski mrežni rječnik – Mrežnik* objavljuje se na stranici rjecnik.hr/mreznik i trenutačno su dostupne natuknice od A do K. Cijeli rječnik (koji se izrađuje u *TLexu*) dostupan je autorici ovoga rada kao suradnici na projektu.



Slika 2. Maksimalna struktura rječničkoga članka u *Mrežniku* (Hudeček i Mihaljević 2021: 22)

4. Dobiveni obrasci

U samoj strukturi rječničkoga članka naglasak je stavljen na rječničke definicije, koje su u prvoj etapi ovoga istraživanja izvezene u poseban dokument kako bi se mogle analizirati te su uočeni jasni obrasci za hiponimiju i meronimiju. Definicije su u najvećoj mjeri sortirane u Pythonu pronalaženjem definicija koje imaju jednaku treću ili četvrtu riječ, ovisno o strukturi definicije. Točnije, u prvome koraku sortirane su definicije prema trećoj riječi tako da se sve definicije koje imaju zajedničku treću riječ pojavljuju jedna ispod druge (tako su dobivene skupine kohiponima, tj. definicija kohiponima). Rezultati su potom pregledani ručno kako bi se provjerilo jesu li obuhvaćene i višerječne leksičke jedinice te su dobivene pravilnosti, tj. leksičko-sintaktički obrasci. Iako pojedini autori ističu da obrasci imaju slab odziv i nisku preciznost (Navigli i Velardi 2010), leksičko-sintaktičke obrasce ipak vrijedi istaknuti jer mogu poslužiti kao osnova za stvaranje razvijenijih modela – posebno obrasce za hrvatski jezik jer dosad nisu opisani.

4.1. Hiponimija

Najveći broj definicija imao je hiponimijsku strukturu, i to tako da se hiponim (podcrtan) izravno definira hiperonomom (podebljan), primjerice:

Agresivnost je **osobina** onoga koji silom djeluje protiv čije volje ili prava.

Bahatost je **osobina** onoga koji je bahat.

Besmrtnost je **osobina** onoga koji je besmrtan, koji živi vječno.

Akcelerator je **uređaj** za povećanje kinetičke energije električno nabijenih čestica.

Alkoholometar je **uređaj** s pomoću kojega se mjeri koncentracija alkohola u organizmu.

Ampermetar je **uređaj** za mjerenje jakosti električne struje.

Najviše je hiponima definirano hiperonimima žena, osobina, svojstvo, uređaj, naprava, mjesto, pripadnik, stanje, predmet, muškarac, znak, boja, prostorija, prostor, način, sustav, riječ, rečenica, razdoblje, zgrada, građevina itd. i hiponimija u takvim definicijama mogla bi se nazvati izravnom hiponimijom. U korpusu osim definicija takva tipa zamjetne su definicije u kojima se između hiponima i hiperonima

pojavljuje i dodatna riječ koja upućuje na to da je riječ o hiponimiji, svojevrsna spona između dviju jedinica koja govori u njihovu značenju i koja je stoga nazvana značenjska spona (usp. Matijević 2024). Kako navode još rani autori poput Crusea (npr. u Cruse 1986: 137), najproznirnija je značenjska spona za hiponimiju vrsta (engl. sort, kind). U Mrežniku se tom riječju u definicijama ustrojavaju brojne hiponimijske hijerarhije, primjerice:

Morseova abeceda vrsta je **telegrafske abecede** koja poruku prikazuje s pomoću točaka i crtica.

Akrostih je vrsta **pjesme** u kojoj početna slova stihova čitana odozgo nadolje daju (...).

Puni pansion vrsta je **ugostiteljske usluge** koja nudi noćenje i tri dnevna obroka.

Osim spone *vrsta* pojavljuju se i spone *rod*, *tip*, *red*, *razred*, *carstvo*, *skupina* – važne u biološkoj sistematici – koje upućuju na hiponimijski odnos. Osim tih spona i na druge se načine upućuje na hiponimiju. Na primjer, izraz *jedan od* također može upućivati na hiponimiju:

Množitelj je jedan od **faktora u računskoj operaciji množenja**, broj kojim se množi.

Svinja je jedan od **znakova kineskoga horoskopa**.

Iako se u radovima o hiponimiji najveći naglasak stavlja na imenice, u analizi je uočena i pravilnost u definicijama glagola:

Cijukati znači **glasati se** glasom koji podsjeća na ciju.

Hukati znači **glasati se** glasom koji podsjeća na hu-hu.

Tuliti znači **glasati se** dugačkim i jednoličnim zvukom.

Hodati znači **kretati se** pješice.

Šetati se znači **kretati se** pješice i polako bez određena cilja radi odmora i razonode.

Trčati znači **kretati se** skačući s noge na nogu.

Klizati se znači **kretati se** po ledu ili po skliskoj površini.

Pritom se kod glagola može uočiti da hiponim i hiperonim imaju jednak glagolski vid, da može postojati i više razina u glagolskoj hijerarhiji (*kretati se* > *hodati* > *šetati se*), ali da i analize glagolskih hiperonima mogu

biti korisne u podjeli glagola na semantičke skupine (navedeni glagoli glasanja i glagoli kretanja).

Za ostale vrste riječi i hiponimiju valja istaknuti da su primjerice brojevi definirani metajezlično, kao *brojevi*, kako su definirani i usklici, čestice i neki veznici te kratice:

Jedan je **broj** koji se obilježuje s 1, za jedan manji od broja dva.

Aha je **usklik** kojim se izriče iznenadno shvaćanje čega.

Npr. je **kratica** od *na primjer*, uvodi primjer.

Definiranje jedne vrste riječi drugom tematiziraju i Lyons (1977: 289–299) i Cruse (1986: 149) te odnos između leksičkih jedinica koje pripadaju različitim vrstama riječi nazivaju kvaziodnosima, tj. navedeni odnos (između npr. *usklik* > *aha*) kvazihiponimijom. U definicijama se mogu primijetiti i druge vrste kvazihiponimije koje autori navode: (1) neslaganje hiponima i hiperonima u broju (npr. *Voće su različiti jestivi, najčešće slatki, plodovi voćaka. – plodovi > voće*) i (2) nepodudaranje u strukturi leksičkih jedinica, kad je hiponim jednorječna leksička jedinica, a hiperonim višerječna ili obratno (npr. *Bacačica diska atletičarka je koja se natječe (...). – atletičarka > bacačica diska*).

U definicijama za hiponimiju važno je primijetiti i definicije u kojima se pojavljuje *istoimen*, među ostalima:

Pauk je pripadnik istoimenoga reda.

Smreka je pripadnica istoimene skupine.

Leksikografi tako upućuju na to da se značenje odnosi na drugo značenje u rječničkom članku i drže se načela da se natuknica ne ponavlja u definiciji, iz čega se zapravo zaključuje da je riječ o hiponimiji unutar jedne leksičke jedinice. Cruse (2000: 110) ističe da jedna riječ može imati općenitije i specifičnije, tj. uže značenje i taj odnos naziva autohiponimijom. U rječničkom korpusu hrvatskoga jezika uočeno je da se ona pojavljuje uglavnom među leksičkim jedinicama koje označuju pripadnike biljnoga i životinjskoga svijeta (npr. *mačke > mačka, smreke > smreka* i dr.).

4.2. Meronimija

U prethodnome poglavlju istaknuto je da se hiponimija često ostvaruje izravno, npr. (*Akcelerator je uređaj (...) – uređaj > akcelerator*), a za meronimiju je najprije važno istaknuti da se ne ostvaruje izravno, nego nužno uz značenjsku sponu. Najčešća je leksička jedinica koja upućuje na meronimiju *dio* (kako je istaknuto u brojnim radovima, npr. još u Winston i dr. 1987: 417–418, Iris i dr. 1988: 269). Najveći broj meronimijskih definicija u *Mrežniku* i ima upravo tu jedinicu (podcrtan meronim, podebljan holonim):

Bat je dio **zvona** koji udaranjem po njegovoj stijenci proizvodi zvuk.
Klip je dio **kukuruza** s malim žutim plodovima u obliku bobica.

Osim jedinice *dio* pojavljuju se jedinice element (*Žila je element lista i stabljike koji služi za (...)*), sastavnica (*Mikroprocesor je sastavnica silikonskoga čipa koja izvršava (...)*), odjel (*Prodaja je odjel u kojoj tvrtki ili djelatnosti koji (...)*) i druge, koje također upućuju na to da je riječ o odnosu dijela i cjeline. Te se jedinice mogu pojaviti i nakon izraza jedan od, koji je spomenut kod hiponimije, pa se pronalazi primjerice:

Struk je jedan od dijelova **sidra**.

Prst je jedan od krajnjih pokretljivih dijelova **šake** i **stopala**.

U definicijama za meronimiju važni su i glagoli, primjerice *sastojati se* (*Jezgra je dio atoma koji se sastoji od protona i neutrona. – atom > jezgra > proton, neutron*), *sadržavati*, *dijeliti se*, ali su za flektivne jezike važni i primjerice *padeži*, tj. sintagmatska razina. Točnije, odnos dijela i cjeline može se izricati pridjevom i imenicom tako da je pridjevom označena cjelina (*rečenični dio – rečenica > dio*) ili dio (sastojak) (*čokoladni kolač – kolač > čokolada*), genitivom *s od* (*ogrtač od balonske svile – ogrtač > balonska svila*), besprijedložnim genitivom (*organ ženskoga tijela – žensko tijelo > organ*), instrumentalom *sa s* (*kolač s nadjevom – kolač > nadjev*), usp. Matijević 2024: 161.

Za meronimiju među glagolima pronađeno je znatno manje primjera nego za hiponimiju, no dva glagola koji stoje kao meronimi (*genus* u definiciji) mogu upućivati na to da radnja koja se definira (natuknica) ima podradnje:

Čitati znači raspoznavati i povezivati slova ili znakove te razumjeti ono što je napisano.

Pušiti znači uvlačiti i ispuštati duhanski dim.

Trepnuti znači nakratko zatvoriti i otvoriti očne kapke.

Među definicijama uočena je kvazimeronimija, najčešće u tome što meronim može biti višerječna leksička jedinica, a holonim jednorječna ili obratno (npr. *Vilica je dio pribora za jelo kojim se nabada hrana. – pribor za jelo > vilica*). Jedinicom *istoimen*, spomenutom kod hiponimije, također se može izraziti da jedna leksička jedinica ima značenje koje označuje cjelinu i jedno koje označuje dio, primjerice:

Badem je jestiv plod gorkaste jezgre istoimene biljke.

Celer je zeleno lišće istoimene biljke koje se upotrebljava kao začim.

Cikla je tamnocrveni jestivi korijen istoimene biljke.

5. Zaključak

Hiponimija i meronimija kao leksičko-semantički odnosi koji ustrojavaju leksičke hijerarhije vrlo su važni za razumijevanje organizacije rječničkoga sustava i mehanizama ljudske kategorizacije. Rad prikazuje kako leksičko-sintaktički obrasci mogu biti dobar putokaz za prikupljanje primjera hiperonima i hiponima te holonima i meronima, koji zatim mogu poslužiti za različita istraživanja, primjerice razradu taksonomije odnosa, uočavanje granica i dodira tih odnosa te za izgradnju računalnih resursa i alata.

Rječnik kao korpus koji se sastoji od strukturiranih i usklađenih definicija može biti dobro polazište za prepoznavanje obrazaca, osobito ako je riječ o računalno čitljivu i dostupnu rječniku koji se može lako pretraživati. Iz *Hrvatskoga mrežnog rječnika – Mrežnika* izvezene su definicije te su uočeni najčešći leksičko-sintaktički obrasci u definicijama. Za hiponimiju su to *HIPONIM je HIPERONIM* (izravna hiponimija) te *HIPONIM je vrsta/rod/tip/red/razred/carstvo/skupina HIPERONIMA*, *HIPONIM je jedan od HIPERONIMA* (gdje postoje značenjske spona između hiponima i hiperonima). Za meronimiju najčešći je obrazac *MERONIM je dio HOLONIMA*, a pojavljuju se i *MERONIM je element/sastavnica/odjel HOLONIMA* ili *MERONIM je jedan od dijelova HOLONIMA*.

Pretraživanje drugih vrsta tekstova, primjerice općih i stručnih korpusa, može dati i nove obrasce – za hiponimiju primjerice *vrsta/tip HIPERONIMA: HIPONIM, (HIPONIM, HIPONIM...), HIPERONIM (HIPONIM,*

HIPONIM...), neki od HIPERONIM je(su) / kao što su: HIPONIM, HIPONIM, HIPONIM (...), HIPERONIM poput HIPONIM i dr. – no detaljnijim proširivanjem analize ovaj bi rad premašio zadani opseg.

Zaključno, istraživanje i dobiveni rezultati (pa tako i primjeri kvaziodnosa i autoodnosa) pokazuju da su rječnici, zahvaljujući strukturi i sustavnosti, plodno tlo za različita jezična istraživanja.

Захвалница: Ovaj rad napisan je na istraživačkome projektu *Hrvatski mrežni rječnik (Mrežnik): 2. faza*, koji financira Europska unija – NextGenerationEU. Za iznesene stavove i mišljenja odgovorna je samo autorica te ti stavovi ne odražavaju službena stajališta Europske unije ili Europske komisije. Ni Europska unija ni Europska komisija ne mogu se smatrati odgovornima za njih.

Literatura

- [1] Ahlswede, T., & Evens, M. (1988). Parsing vs. text processing in the analysis of dictionary definitions. U: *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics* (pp. 217–224). Association for Computational Linguistics. <https://doi.org/10.3115/982023.982050>
- [2] Amsler, R. (1980). *The structure of the Merriam-Webster pocket dictionary*. University of Texas.
- [3] Ando, M., Sekine, S., & Ishizaki, S. (2004). Automatic extraction of hyponyms from Japanese newspapers using lexico-syntactic patterns. U: M. T. Lino, M. F. Xavier, F. Ferreira, R. Costa, & R. Silva (ur.), *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)* (pp. 387–390). European Language Resources Association.
- [4] Baisa, V., & Suchomel, V. (2015). Corpus-based extraction of hypernyms in terminological thesaurus for land surveying domain. U: *Ninth Workshop on Recent Advances in Slavonic Natural Language Processing* (pp. 69–74). Tribun EU.
- [5] Barbu Mititelu, V. (2011). Hyponymy patterns in Romanian. *Memoirs of the Scientific Sections of the Romanian Academy*, 34, 1–13.
- [6] Barnbrook, G. (2002). *Defining language: A local grammar of definition sentences*. *Studies in Corpus Linguistics*. John Benjamins Publishing Company.
- [7] Berland, M., & Charniak, E. (1999). Finding parts in very large corpora. U: *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL 1999)* (pp. 57–64).
- [8] Bielinskienė, A., Boizou, L., Kovalevskaitė, J., & Utkā, A. (2012). Towards the automatic extraction of term-defining contexts in Lithuanian. U: A. Tavast, K.

- Muischnek, & M. Koit (ur.), *Human language technologies: The Baltic perspective* (pp. 18–26). IOS Press.
- [9] Cruse, D. A. (1986). *Lexical semantics*. Cambridge University Press.
- [10] Cruse, D. A. (2000). *Meaning in language: An introduction to semantics and pragmatics*. Oxford University Press.
- [11] Čulig Suknaić, J. (2020). *Antonimija kao pojmovna kategorija značenjske suprotnosti u engleskome i hrvatskome jeziku*. University of Zagreb, Faculty of Humanities and Social Sciences. <https://doi.org/10.17234/diss.2020.201018>
- [12] Farrugia, J. P. (2019). *Exploring hypernym discovery*. University of Malta, Department of Artificial Intelligence.
- [13] Fellbaum, C. (1998). *WordNet – An electronic lexical database*. MIT Press.
- [14] Geeraerts, D. (1999). Hundred years of lexical semantics. U: M. Vilela & F. Silva (ur.), *Atas do 1º Encontro internacional de linguística cognitiva* (pp. 123–154). Faculdade de Letras.
- [15] Geeraerts, D. (2010). *Theories of lexical semantics*. Oxford University Press.
- [16] Gerstl, P., & Pribbenow, S. (1996). A conceptual theory of part-whole relations and its applications. *Data & Knowledge Engineering*, 20(3), 305–322.
- [17] Gil-Berrozpe, J. C. (2017). *Corpus-based identification of hyponymy subtypes and knowledge patterns in the environmental domain*. University of Granada, Faculty of Translation and Interpreting.
- [18] Gil-Berrozpe, J. C., León-Araúz, P., & Faber, P. (2017). Specifying hyponymy subtypes and knowledge patterns: A corpus-based study. U: I. Kosem, C. Tiberius, M. Jakubiček, J. Kallas, S. Krek, & V. Baisa (ur.), *Electronic lexicography in the 21st century: Proceedings of eLex 2017 conference* (pp. 63–92). Lexical Computing CZ s.r.o.
- [19] Girju, R., Badulescu, A., & Moldovan, D. (2006). Automatic discovery of part-whole relations. *Computational Linguistics*, 32(1), 83–135.
- [20] Hearst, M. (1992). Acquisition of hyponyms from large text corpora. U: *Proceedings of the 14th International Conference on Computational Linguistics (COLING-92)* (pp. 539–545). Nantes.
- [21] Hudeček, L., & Mihaljević, M. (2021). *Hrvatski mrežni rječnik – MREŽNIK: Monografija (radna inačica)*. Institut za hrvatski jezik i jezikoslovlje.
- [22] Hudeček, L., Mihaljević, M., & Jozić, Ž. (ur.). (2024). *Anatomija rječnika: Hrvatski mrežni rječnik – Mrežnik*. Institut za hrvatski jezik.
- [23] Iris, M. A., Litowitz, B. E., & Evens, M. (1988). Problems of the part-whole relation. U M. W. Evens (ur.), *Relational models of the lexicon: Representing knowledge in semantic networks* (pp. 261–288). Cambridge University Press.

- [24] Jones, S. (2006). A lexico-syntactic analysis of antonym co-occurrence in spoken English. *Text & Talk*, 26(2), 191–216.
- [25] Jones, S. (2010). Using web data to explore lexico-semantic relations. U: P. Starjohann (ur.), *Lexical-semantic relations: Theoretical and practical perspectives*. John Benjamins.
- [26] Jones, S., & Murphy, L. M. (2005). Using corpora to investigate antonym acquisition. *International Journal of Corpus Linguistics*, 10(3), 401–422.
- [27] Justeson, J. S., & Katz, S. M. (1991). Co-occurrences of antonymous adjectives and their contexts. *Computational Linguistics*, 17(1), 1–19.
- [28] Liu, L., Cao, C., & Wang, H. (2006). Extracting hyponymic relations from Chinese free corpus. U: *Proceedings of the 5th WSEAS International Conference on Applied Computer Science (ACOS'06)* (pp. 962–968). WSEAS.
- [29] Lyons, J. (1963). *Structural semantics: An analysis of part of the vocabulary of Plato*. Basil Blackwell.
- [30] Lyons, J. (1977). *Semantics* (Vols. 1–2). Cambridge University Press.
- [31] Markowitz, J., Ahlswede, T., & Evens, M. (1986). Semantically significant patterns in dictionary definitions. U: *Proceedings of the 24th Annual Conference of the ACL* (pp. 112–119).
- [32] Materynska, O. (2019). Semantic relation of meronymy in languages of different structure (case study of semantics of body part names). *Advanced Education*, 12, 256–264.
- [33] Matijević, M. (2024). *Hijerarhijski leksičko-semantički odnosi u hrvatskome jeziku*. Filozofski fakultet Sveučilišta u Zagrebu.
<https://doi.org/10.17234/diss.2025.314038>
- [34] Miller, G. A. (1990). Nouns in WordNet: A lexical inheritance system. *International Journal of Lexicography*, 3(4), 246–264.
- [35] Navigli, R., & Velardi, P. (2010). Learning word-class lattices for definition and hypernym extraction. U: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (pp. 1318–1327).
- [36] Petrović, B. (2005). *Sinonimija i sinonimičnost u hrvatskome jeziku*. Hrvatska sveučilišna naklada.
- [37] Przepiórkowski, A., Degórski, Ł., Spousta, M., Simov, K., Osenova, P., Lemnitzer, L., Kuboň, V., & Wójtowicz, B. (2007). Towards the automatic extraction of definitions in Slavic. U: *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing 2007: Special theme – Information extraction and enabling technologies* (pp. 43–50). Association for Computational Linguistics.

- [38] Rydin, S. (2002). Building a hyponymy lexicon with hierarchical structure. U: *Proceedings of the ACL-02 Workshop on Unsupervised Lexical Acquisition* (pp. 26–33).
- [39] Sahin, G. (2017). Extraction of hyponymy, meronymy, and antonymy relation pairs: A brief survey. *International Journal on Natural Language Computing*, 6(2), 1–18.
- [40] Sang, E. T. K., & Hofmann, K. (2007). Automatic extraction of Dutch hypernym-hyponym pairs. *LOT Occasional Series*, 7, 163–174.
- [41] Şerbetçi, A., Orhan, Z., & Pehlivan, İ. (2011). Extraction of semantic word relations in Turkish from dictionary definitions. U: *Proceedings of the ACL 2011 Workshop on Relational Models of Semantics* (pp. 11–18). Association for Computational Linguistics.
- [42] Snow, R., Jurafsky, D., & Ng, A. (2005). Learning syntactic patterns for automatic hypernym discovery. U: *Advances in neural information processing systems* (pp. 1297–1304).
- [43] Stanković, R., Krstev, C., Stijović, R., Gočanin, M., & Škorić, M. (2021). Towards automatic definition extraction for Serbian. U: L. Mitits & S. Kiosses (ur.), *Proceedings of the XIX EURALEX Congress: Lexicography for Inclusion* (Vol. 2, pp. 695–703). Democritus University of Thrace.
- [44] Šarić, Lj. (2007). *Antonimija u hrvatskome jeziku: Semantički, tvorbeni i sintaktički opis*. Hrvatska sveučilišna naklada.
- [45] Šarić, Lj., & Wittschen, W. (2008). *Rječnik sinonima hrvatskoga jezika*. Jesenski i Turk.
- [46] Vintar, Š., & Gorjanc, V. (2003). Identifying markers of semantic relations in Slovene. *Strani jezici*, 32(1–2), 37–44.
- [47] Winston, M. E., Chaffin, R., & Herrmann, D. (1987). A taxonomy of part-whole relations. *Cognitive Science*, 11(4), 417–444.
- [48] Yıldırım, S., & Yıldız, T. (2012a). Corpus-driven hyponym acquisition for Turkish language. U: A. Gebukh (ur.), *Computational linguistics and intelligent text processing (CICLing 2012)* (pp. 29–41). Springer.
- [49] Yıldırım, S., & Yıldız, T. (2012b). Automatic extraction of Turkish hypernym-hyponym pairs from large corpus. U: *Proceedings of the International Conference on Computational Linguistics (COLING 2012)* (pp. 493–500).
- [50] Yıldız, T., Diri, B., & Yıldırım, S. (2016). Acquisition of Turkish meronym based on classification of patterns. *Pattern Analysis and Applications*, 19(2), 495–507.
- [51] Zgusta, L. (1971). *Manual of lexicography*. De Gruyter Mouton.

From a Dictionary as Corpus to Hyponymy and Meronymy

Maja Matijević

Summary

This study investigates two hierarchical lexical-semantic relations, hyponymy and meronymy, and their semi-automatic identification in dictionary definitions. Hyponymy is defined as the relation between a lexical unit denoting a category and a lexical unit denoting a member of that category (e.g., hyperonym *animal* > hyponym *dog*), while meronymy is defined as the relation between a lexical unit denoting a whole and a lexical unit denoting a part of that whole (e.g., holonym *body* > meronym *hand*). The study establishes a theoretical framework and demonstrates the potential of automated identification in dictionary definitions. The methodological approach is based on lexico-syntactic patterns, i.e., typical structures characterized by specific syntactic regularities and lexical units that enable the recognition of lexical-semantic relations. The analysis of dictionary definitions yielded several patterns, with particular emphasis on case forms and prepositional expressions relevant for expressing part-whole relations. The results highlight dictionaries as a rich and reliable source for the recognition of hierarchical relations, opening new perspectives for further linguistic research and analysis.

Keywords: lexico-semantic relations, hyponymy, meronymy, lexico-syntactic patterns, dictionary

Metodologija izrade višejezičnog paralelnog korpusa na osnovu onlajn digitalnih uputstava za upotrebu: korpus Hilti uputstava

Naučni rad

DOI: 10.18485/judig.2025.1.ch29

Nikola Janković¹,  0000-0003-3484-4220

Apstrakt

U radu je predstavljena metodologija izrade opsežnog višejezičnog paralelnog korpusa uputstava za upotrebu na osnovu digitalnih uputstava za upotrebu uređaja kompanije Hilti. Cilj ovog rada je doprinos boljim resursima za istraživanja vezanim za mašinsko prevođenje, teoriju prevođenja, ali i za nastavu prevođenja i druge vrste lingvističkih istraživanja. Polazna tačka metodologije bilo je korišćenje uniformne strukture HTML uputstava za svaki jezik na zvaničnoj Hilti onlajn bazi uputstava. Dalji koraci su obuhvatali identifikovanje HTML elemenata koji sadrže relevantne delove teksta, njihovu doslednu numeraciju za svaki jezik, identifikovanje i čišćenje pogrešno paralelizovanih tekstova i čuvanje obrađenih tekstova u formatu TMX, tako da je engleski izvorni jezik u svakoj datoteci. U proseku, za svako uputstvo postoji oko 20 ciljnih jezika, a sam korpus sadrži oko 7 miliona reči. Prednosti kreiranog korpusa uključuju visoku preciznost kada je u pitanju podudaranje segmenata između jezika, značajan broj reči, kao i veliki broj jezika obuhvaćenih korpusom. U radu će biti detaljno opisani koraci preuzimanja, obrade i strukturiranja podataka, tehnički izazovi sa kojima smo se suočili i kako su rešeni, kao i osnovni statistički podaci vezani za jezike u korpusu. Smatramo da se ovakav tip korpusa može koristiti za unapređivanje resursa za prevođenje, teoriju prevođenja, nastavu prevođenja, kao i za različite vrste analiza stručnog jezika.

Ključne reči: paralelni korpusi, mašinsko prevođenje, jezički alati, stručni jezik, TMX

¹ Filološki fakultet, Univerzitet u Beogradu, nikolajankovickv@gmail.com

1. Uvod

Pod terminom „paralelni korpusi“ najčešće se podrazumevaju korpusi koji sadrže tekstove na jednom prirodnom jeziku (takozvanom „izvornom jeziku“), uparene sa njihovim prevodom na drugom, takozvanom „ciljnom“ jeziku, iako ovaj termin može obuhvatati i korpuse tekstova na različitim jezicima koji su uporedivi na osnovu određenih kriterijuma poput oblasti znanja (Lefer, 2020:257-258). Osim toga, moguće je da paralelni korpus sadrži dva različita prevoda na isti jezik i u tom slučaju se izvorni i ciljni jezik poklapaju. Posle njihove pojave tokom 1990-ih godina prošlog veka, paralelni korpusi su postali vredan resurs u oblastima kontrastivne lingvistike, bilingvalne leksikografije, nastavi stranih jezika, nastavi prevođenja, ekstrakciji termina, mašinskom prevođenju i računarski potpomognutom prevođenju (engl. computer-aided translation), učenju parafraza, projekciji jezičkih alata na druge jezike, kao i mnogim drugim primenama u obradi prirodnih jezika (Bañón et al., 2020:4555; Lefer, 2020:258).

Međutim, za izradu ove vrste korpusa vezuju se specifične prepreke. Kako navode Lefer i saradnici (Lefer, 2020:268), prvi problem vezan za izradu paralelnih korpusa je problem dostupnosti tekstova. Naime, za određene jezičke parove postoji veoma mali broj paralelnih tekstova, pri čemu tipovi teksta koji se često prevode mogu biti teško dostupni zbog faktora poverljivosti (ibid.). Takođe, kako Varga i saradnici (Varga et al., 2007:247) navode, 40% svetske populacije su maternji govornici jezika sa 100 miliona govornika ili više, dok 4% svetske populacije su maternji govornici jezika sa pola miliona govornika ili manje; najveći broj ljudi na svetu govori takozvane „jezike srednje gustine“ (engl. *medium density languages*), čiji broj maternjih govornika je između ova dva broja. Međutim, kako isti autori ističu, broj paralelnih strana na Internetu koji uključuje jezike srednje gustine je iznenađujuće mali (ibid:247). Sa druge strane, sastavljanje paralelnih korpusa iz tekstualnih izvora koji nisu nastali u elektronskom (engl. *digital-born*) formatu podrazumeva optičko prepoznavanje teksta (engl. *optical character recognition*, skr. OCR) koje značajno otežava proces obrade tekstova zbog neizbežnih grešaka koje uvodi (Lopresti, 2009), dok paralelni dokumenti u elektronskom formatu u specijalizovanim domenima, kao što je pomenuto, često nisu dostupni zbog pitanja poverljivosti. Takođe, jedan od najzahtevnijih koraka u kreiranju paralelnih korpusa predstavlja segmentacija tekstova i njihovo uparivanje (na nivou paragrafa, rečenice ili reči), što najčešće zahteva korak ručne provere. Iz navedenih razloga smatramo da sajtovi sa veb-stranama na različitim jezicima, koje imaju identičnu HTML strukturu, predstavljaju

veoma vredan resurs za efikasnu izradu kvalitetnih paralelnih korpusa. U ovom radu predstavljen je Hilti-UM korpus (prema engleskoj sintagmi *Hilti User Manuals*), napravljen na osnovu onlajn uputstava za upotrebu za uređaje kompanije Hilti, a čija metodologija izrade može biti korisna osnova za izradu sličnih korpusa, naročito u oblasti stručnih jezika.

2. Proces izrade Hilti-UM korpusa

U fazi istraživanja izvora za građu paralelnih korpusa stručnog jezika koji bi uključivali srpski jezik, početna hipoteza je bila da pogodan format predstavljaju višejezična uputstva za upotrebu u mašinski čitljivom formatu, konkretno PDF. Međutim, iz razloga navedenih u odeljku 5, veb-strane dostupne na više jezika koje imaju identičnu HTML strukturu su se pokazale kao znatno jednostavniji format za obradu, a sajt kompanije Hilti kao najbolji kandidat za izvor građe takve vrste.

Prvi korak u procesu sakupljanja građe predstavljao je pretragu dokumenata na sajtu kompanije Hilti² iz kategorija "uputstvo za rukovanje" i "uputstvo za upotrebu", uz filter "srpski jezik".

U narednom koraku je sastavljena lista uputstava dostupnih u formatu HTML pomoću *Python* skripte i modula *Selenium* za automatizaciju pretraživača, pomoću kog su preuzeti i sačuvani u formatu JSON (*lista_uputstava.json*) linkovi za svaki model proizvoda, nazivi modela, kao i jezici za koje je uputstvo dostupno. Na osnovu sačuvanih linkova, za određeni model proizvoda su automatski preuzeta HTML uputstva na svim dostupnim jezicima pomoću *Python* skripte (*preuzimanje.py*) koja generiše odgovarajuće linkove za preuzimanje. *Python* skripta koristi prethodno napravljen *Python* rečnik kojim se uparuju nazivi jezika na sajtu sa njihovim jezičkim kodom (npr. "nemački": "de", "norveški": "no", "finski": "fi", "ruski": "ru") kako bi dodavanjem odgovarajućih skraćenica za jezik automatski bili generisani linkovi ka uputstvima na svim dostupnim jezicima. Za svaki model proizvoda je na njegov osnovni URL (na primer, <https://www.hilti.rs/content/hilti/E4/RS/sr/op-man.html/2201898>) dodat svaki odgovarajući jezički kod (npr. "/en", "/sr", "bg"), a sadržaj ovih strana je sačuvan u zasebnom katalogu imenovanom kao i sam model. Obrazac imenovanja datoteka bio je *{naziv modela}_{jezički kod}.htm*, pri čemu su eventualne kose crte (/ i \) zamenjene crticama (-) kako bi nazivi datoteka bili validni. U ovoj fazi pronađeno je 120 modela koji poseduju HTML verziju uputstava, uspešno su preuzete HTML datoteke za 112 modela od

² <https://www.hilti.rs/downloads.html>

kojih je 108 imalo uputstvo na engleskom jeziku, odabranom za izvorni jezik.

Pomoću druge Python skripte (*paralelizacija.py*), koristeći modul *Beautiful Soup*, obrađena je sadržina svih HTML datoteka. Python skripta iz odgovarajućeg odeljka (HTML elementa čija klasa ima vrednost *o-editorial-section*, a nalazi se u elementu *main*) ekstrahuje sve HTML elemente koji pripadaju sledećim tipovima: *p*, *h1*, *h2*, *h3*, i *li*. Tekstualni sadržaj ekstrahovanih elemenata (koji odgovaraju pasusima) u izvornoj i ciljnoj datoteci potom se zapisuje u *Python* liste. S obzirom na to da je struktura HTML datoteka identična za sve jezike u okviru jednog modela, elementi tako dobijenih *Python* lista se zatim mogu jednostavno upariti,.

Iz prethodno sačuvane *JSON* datoteke (*lista_uputstava.json*) sa osnovnim podacima vezanim za uputstva, prema prethodno utvrđenom šablonu automatski je generisan URL za englesku verziju HTML uputstva (koja je uzeta kao izvorni tekst), i za verziju na drugom jeziku (koja je uzeta kao ciljni tekst).

U finalnoj fazi paralelizacije generiše se *TMX* datoteka u kojoj su upareni pasusi iz engleskog i odgovarajućeg ciljnog jezika. Jedan par pasusa se nalazi u elementu *tu* (engl. translation unit = jedinica prevođenja), koji sadrži po jedan element *tuv* za izvorni jezik (engleski, sa atributom *xml:lang="en"*) i ciljni jezik sa odgovarajućom vrednošću atributa *xml:lang*. Ovaj par elemenata *tuv* sadrži upareni tekst originalnih pasusa iz izvornog i ciljnog jezika. *TXM* datoteke se čuvaju u potkatalogu *txm* odgovarajućeg kataloga modela.

3. Validacija paralelizacije

Po obavljenoj paralelizaciji pasusa, odnosno generisanju *TMX* datoteka, bilo je potrebno izvršiti validaciju njenog kvaliteta za svaki model i svaki jezički par. S obzirom na veliki broj modela, jezičkih parova i pasusa u svakom dokumentu, ručna provera paralelizacije izvršena je za sve srpsko-engleske parove, dok je za ostale jezičke parove validacija izvršena programskim putem u dve faze pomoću dve *Python* skripte (*validacija.py* i *uzastopni_loši_segmenti.py*).

Prva faza programske validacije podrazumeva validaciju uparenih pasusa (obuhvaćenih elementima *tu*) na osnovu nekoliko kriterijuma (koristeći fajl *validacija.py*). Validan paralelizovan par je morao da zadovolji sledeće kriterijume:

- Postoji element *tuv* i za izvorni i za ciljni jezik, i nijedan od njih nije prazan.

- Ukoliko postoje brojevi u izvornom elementu, identični brojevi moraju postojati i u ciljnom elementu.
- Ukoliko se u izvornom elementu nalazi neki od sledećih simbola: %, ©, ®, ™, \$, €, £, ili ¥, isti simbol mora postojati i u ciljnom elementu.
- Ukoliko je tekst u izvornom elementu duži od 10 reči, razlika u dužini izvornog i ciljnog teksta ne sme biti veća od 50%. Ovaj odnos izračunat je pomoću sledećeg izraza: $\frac{\min(\text{len}(\text{source\_text}), \text{len}(\text{target\_text}))}{\max(\text{len}(\text{source\_text}), \text{len}(\text{target\_text}))}$.

Ova provera obuhvatala je jezike koji koriste ćirilčno ili latinično pismo. Nije bilo moguće automatski validirati hebrejski, japanski, jermenski, korejski, kineski i kazaški jezik, i samim tim je provera TMX datoteka, kojima je jedan od navedenih jezika bio ciljni, automatski preskočena i one su premeštene u zaseban katalog.

Rezultati ove analize sačuvani su u JSON datoteku (*rezultati_validacije.json*), uključujući podatke koje su datoteke (ne)obrađene, a za svaku obrađenu datoteku i informacije o elementima koji su prekršili neki od prethodno navedenih kriterijuma validnog paralelizovanog para. Informacije o ovakvim elementima obuhvataju identifikacioni broj elementa, kriterijum koji je prekršen, izvorni tekst i ciljni tekst i one pružaju neophodni kontekst za dalju doradu kriterijuma, kako bi se obuhvatili svi nepravilno paralelizovani tekstovi i smanjio broj lažno pozitivnih rezultata.

JSON datoteka sa rezultatima analize paralelizacije (*rezultati_validacije.json*), proistekla iz prethodne faze validacije paralelizacije, iskorišćena je u potonjoj fazi validacije. Na osnovu nekoliko ponovljenih testiranja različitih kriterijuma, najbolji rezultati prilikom filtriranja pogrešno paralelizovanih datoteka postignuti su primenom kriterijuma uzastopnih pogrešnih elemenata. Naime, pomoću *Python* skripte *uzastopni_loši_segmenti.py*, datoteke koje su sadržale 5 ili više uzastopnih pogrešnih elemenata su automatski označene kao pogrešno paralelizovane. Na osnovu ovog kriterijuma, kreirana je finalna JSON datoteka (*finalna_validacija.json*) sa listom nepravilno paralelizovanih datoteka, kao i listom datoteka čija je obrada automatski preskočena i koje su zatim premeštene u zasebne kataloge. Finalni validirani korpus obuhvatio je 54 različita modela sa ukupno 800 TMX datoteka.

4. Analiza korpusa Hilti-UM

Posle faze izrade TMX datoteka i njihove validacije, pomoću *Python* skripte (*analiza_korpusa.py*) urađeno je nekoliko jednostavnih vrsta analiza vezanih za jezik sadržan u samom korpusu.

Glavni katalog korpusa sadrži kataloge pojedinačnih modela mašina čija su uputstva paralelizovana. Obilaskom potkataloga glavnog kataloga korpusa pomoću odgovarajuće *Python* skripte, kreiran je indeks TMX datoteka, gde je za svaku TMX datoteku zabeležen ciljni jezik na osnovu atributa *xml:lang* elemenata *tuv* (engleski jezik je bio izvorni jezik u svim datotekama).

Po kreiranju indeksa TMX datoteka izvršena je analiza sadržaja samih datoteka na nivou tokena (posebno reči) i segmenata. Kada je u pitanju analiza na nivou reči, s obzirom na veliki broj jezika koji je zastupljen u korpusu, bilo je potrebno uzeti u obzir specifičnosti kodiranja različitih pisama i njihovih karaktera kako bi se poboljšala tačnost tokenizacije, odnosno segmentacije teksta na reči. Na osnovu smernica navedenih u aneksu Unikod konzorcijuma o segmentaciji teksta (Unicode Consortium, 2025) za segmentaciju na nivou reči je primenjen Unikod regularni izraz:

$\backslash p\{L\}[\backslash p\{L\}\backslash p\{M\}\backslash p\{Nd\}\backslash p\{Pc}\"] *$ (RI 1)

Regularni izraz (RI 1) pronalazi tekst koji odgovara sledećim pravilima:

- Počinje slovom ($\backslash p\{L\}$) posle kojeg ne sledi nijedan ili sledi bar jedan (*) karakter koji može biti:
 - slovo ($\backslash p\{L\}$),
 - znak interpunkcije za povezivanje (engl. *combining diacritics*) ($\backslash p\{M\}$),
 - cifra³ ($\backslash p\{Nd\}$),
 - znak vezne interpunkcije (engl. *connector punctuation*) ($\backslash p\{Pc\}$), poput crtice (-),
 - apostrof (').

Pomoću *Python* biblioteke *lxml* (konkretno, klase *xml.etree.ElementTree* i funkcije *iterparse*), obrađeni su svi segmenti u

³ S obzirom na tehničku prirodu tekstova u korpusu Hilti-UM, određeni broj termina i oznaka sadrži i slova i brojeve, zbog čega je ovaj skup karaktera dodat kao moguć sastavni deo unutar reči. Cifre na početku nisu uključene u reč, čime se izbegava da se mere, poput "8mm", beleže kao reči.

TMX datotekama pojedinačno, beležeći pritom broj segmenata, broj reči, kao i same reči koje se pojavljuju za određeni jezik. S obzirom na to da je engleski jezik bio izvorni jezik u svim datotekama, ovi podaci su se za engleski jezik sakupljali na osnovu podataka za izvorni jezik samo jedne TMX datoteke za svaki od modela, te broj dokumenata za engleski jezik odgovara broju modela čija su uputstva paralelizovana. Analiza je pokazala da validiran korpus obuhvata 3.439.354 tokena u okviru 800 datoteka, u kojima se pojavljuje 26 jezika.

Za svaki jezik, izračunat je broj dokumenata u kojima je jezik korišćen, broj segmenata, broj tokena, broj jedinstvenih reči i prosečan broj reči po segmentu (Tabele 1.1 i 1.2).

S obzirom na to da je u fazi pretrage uputstava na sajtu firme Hilti primenjen filter kojim se prikazuju samo uputstva koja postoje i na engleskom i na srpskom jeziku, ova dva jezika su zastupljena u najvećem broju modela (54 za engleski i 43 za srpski jezik), pri čemu je razlika u broju modela (54 – 43) posledica toga što 9 TMX datoteka na srpskom jeziku nije prošlo fazu validacije paralelizacije. Izuzimajući jezike isključene iz ove verzije korpusa zbog nemogućnosti validacije paralelizacije, kao i jezike za koje je validacija paralelizacije bila (gotovo) uvek neuspešna (zbog drugačijeg rasporeda sadržaja dokumenta), jezici sa najmanje pojavljivanja u modelima su danski, estonski i norveški (18 TMX datoteka). Prosečan broj dokumenata u kojima se koristi određeni jezik je 32,8. Kada je u pitanju broj tokena po jezicima, najveći broj tokena je očekivano prisutan u engleskom jeziku, dok je broj tokena za srpski potkorpus manji od onog za španski, francuski, italijanski, portugalski, rumunski i grčki uprkos većem broju dokumenata sa srpskim jezikom.

Tabela 1.1: Broj tokena po jezicima

Jezik	Broj tokena
en	236006
es	210535
fr	191069
it	190065
pt	186244
ro	185904
el	182446
sr	175374

Tabela 1.2: Broj dokumenata po jezicima

Jezik	Broj dokumenata
en	54
sr	43
es	42
hr	42
ro	41
it	40
nl	40
pt	40

nl	163977	bg	39
bg	162414	el	39
hr	161062	fr	39
sl	151499	sk	37
pl	137287	sl	37
sk	135374	cs	36
cs	132637	pl	35
uk	121691	de	31
de	117787	hu	31
hu	116690	uk	28
tr	93104	tr	26
sv	64986	lt	21
lt	64923	lv	20
lv	62343	sv	20
no	59817	fi	19
da	59404	da	18
fi	51414	et	18
et	50134	no	18
prosečno	133237.9	prosečno	32.8

Koristan uvid u leksiku u specijalizovanom domenu, kao što su uputstva za upotrebu različitih vrsta aparata, predstavljaju frekvencijske liste. Kako bismo izdvojili najčešće termine u engleskom i srpskom jeziku koji su česti u ovoj vrsti tekstova, ali ne i u opštem jeziku, koristili smo liste frekvencija generisane na osnovu velikih korpusa engleskog, odnosno srpskog jezika. Za engleski jezik korišćena je lista najfrekventnijih 333.333 reči iz Guglovog korpusa (en. *Google corpus*) (Norvig, 2009), dok je za srpski jezik korišćena lista frekvencija od preko 2 miliona srpskih reči, dobijena na osnovu srpskih potkorpusa velikih veb-korpusa i velikih elektronskih korpusa srpskog jezika⁴, među kojima izdvajamo korpuse Kišobran i Znanje (Škorić & Janković, 2024). U nedostatku specijalizovane liste frekvencija učestalih reči koje imaju malu informativnu vrednost (engl. *stopwords*) za srpski jezik, koristili smo jednostavno filtriranje najčešćih 1241 reči iz navedenih lista frekvencija. Ovaj broj je odabran na

⁴ <https://huggingface.co/datasets/procesaur/Wikipedija>,
<https://huggingface.co/datasets/procesaur/kisobran>,
<https://huggingface.co/datasets/procesaur/znanje>, <https://huggingface.co/datasets/oscar-corpus/oscar>

osnovu istraživanja u kome su Marovac i saradnici (Marovac et al., 2021) izdvojili 1241 najučestaliju reč male informativne vrednosti u srpskom jeziku. Pritom ne odbacujemo mogućnost da su neki validni termini isključeni iz liste zbog jednostavnosti naše metodologije. Tabele 2.1 i 2.2 ilustruju liste najčešćih termina za srpski i engleski jezik u korpusu Hilti-UM.

Tabela 2.1: Lista najčešćih termina u srpskom potkorpusu

	reč	frekvencija
1	baterije	1927
2	akumulatorske	1600
3	alata	1197
4	alat	1066
5	bateriju	1027
6	akumulatorsku	953
7	baterija	921
8	koristite	712
9	nemojte	677
10	hilti	656
11	proizvod	650
12	električni	629
13	električnog	589
14	back	552
15	akumulatorskih	519

Tabela 2.2: Lista najčešćih termina u engleskom potkorpusu

	reč	frekvencija
1	battery	3528
2	tool	3196
3	batteries	1836
4	instructions	1213
5	hilti	889
6	switch	841
7	operating	827
8	accessory	797
9	dust	749
10	clean	659
11	remove	629
12	injury	579
13	damage	552
14	operation	550
15	damaged	522

Korpus Hilti-UM će biti dodat na platformu „Bibliša“⁵ kako bi bio dostupan u formatu pogodnom za jezičke analize i upotrebu u prevodilačkoj praksi.

5. Prepreke pri izradi korpusa Hilti-UM

U ovom odeljku predstavimo glavne prepreke sa kojima smo se susreli u izradi korpusa Hilti-UM, kao i rešenja koja su primenjena za njihovo prevazilaženje.

Najpre, kada je u pitanju faza sakupljanja građe za paralelni korpus, uputstva za upotrebu različitih aparata se ubrajaju u lakše dostupne resurse stručnih paralelnih tekstova na različitim jezicima, uključujući srpski. Prva prepreka, sa kojom smo se susreli prilikom pretrage sajtova različitih proizvođača raznih vrsta proizvoda (mašina, aparata) na srpskom tržištu i

⁵ <http://biblisha.jerteh.rs/>

analize njihove dostupnosti, je format uputstava. Naime, najveći broj pronađenih uputstava, dostupnih onlajn, nalazi se u formatu PDF koji nije mašinski čitljiv. Otuda je neophodan proces optičkog prepoznavanja karaktera (račitavanja teksta) u uputstvima, što predstavlja veliki izazov u višejezičnim dokumentima ove vrste. Takođe, prisustvo slikovnih uputstava u PDF dokumentima uslovljava da redosled istih pasusa na različitim jezicima u određenim slučajevima bude različit, što za posledicu zahteva dugotrajan proces ručnog uparivanja pasusa. Imajući u vidu ova ograničenja, datoteke u formatu PDF nisu tretirane kao poželjni izvor građe za ovakav korpus.

U nastavku istraživanja sajtova proizvođača mašina, ustanovili smo da neki od njih nude uputstva za upotrebu u formatu HTML, kao posebne stranice svog sajta. Među ovim firmama, sajt kompanije Hilti je u ponudi imao najveći broj višejezičnih uputstava za različite modele u formatu HTML. Ručnim upoređivanjem HTML strukture uputstava na različitim jezicima i za različite modele, utvrdili smo da je struktura uputstava u najvećem broju slučajeva istovetna, što je u velikoj meri omogućilo automatsko sakupljanje građe opisano u odeljku 2.

U ovoj fazi, bilo je važno razmotriti pravne i etičke faktore sakupljanja građe na Internetu u svrhe izrade korpusa, naročito kada je u pitanju automatsko sakupljanje građe sa veb-sajtova (engl. *web scraping*). Naime, kako navode Krotov i saradnici (Krotov et al., 2020:558-559), osim poznavanja alata za preuzimanje i obradu podataka sa Interneta, istraživači moraju imati u vidu i načine upotrebe ovih podataka na zakonit i etičan način i preduzmu proaktivne mere da osiguraju da je njihovo sakupljanje podataka na Internetu u skladu sa pravnim i etičkim standardima. U ovom radu, autori su analizirali 147 naučnih radova vezanih za tehničke, etičke, i pravne aspekte preuzimanja podataka sa Interneta za naučne svrhe i predložili Etički i pravni okvir za automatsko preuzimanje podataka sa Interneta (en. *Legality and Ethics Framework for Web Scraping*) koji istraživači mogu pratiti pri donošenju odluka o (automatskom) sakupljanju podataka sa Interneta i koji obuhvata različite kriterijume u formi pitanja koja se tiču ilegalnog pristupa i korišćenja podataka, kršenja ugovora, prouzrokovanja štete i poslovnih tajni. Izrada korpusa Hilti-UM je vršena u skladu sa navedenim okvirom, zadovoljavajući pritom sve navedene kriterijume u ovom okviru. Faktori koji su omogućili da, prema našem istraživanju, ovi kriterijumi budu zadovoljeni su sledeći:

- Protokol *robots.txt* sajta *hilti.rs*, koji reguliše aktivnosti automatskog pretraživanja i preuzimanja na sajtu, dozvoljava pristup stranici za preuzimanje uputstava za upotrebu,

- Sama uputstva su javno dostupna i namenjena širokoj divulgaciji,
- Odeljak „Uslovi i politika“ sajta *hilti.rs*, koji se tiče prodaje i pružanja usluga Hilti proizvoda, ne pominje sadržaj na samom sajtu niti njegovo preuzimanje,
- Osim autorskih prava vezanih za prodate proizvode u dokumentu „Uslovi i politika“, kao i napomene vezane za Hilti zaštitni znak u uputstvima za upotrebu, u našem istraživanju samog sajta, dokumenta „Uslovi i politika“, kao i samih uputstava nismo našli eksplicitno naglašavanje autorskih prava o podacima na sajtu,
- Automatsko preuzimanje uputstava sa sajta Hilti je vršeno uz programirano pauziranje između zahteva kako se ne bi vršilo opterećenje servera,
- Proizvod ovog istraživanja je nekomercijalni paralelni korpus namenjen naučnom istraživanju, koji će biti prezentovan u vidu izdvojenih paralelnih segmenata, pomoću kojih se ne može rekonstruisati originalni tekst.

Kao što Krotov i saradnici (ibid.:567) naglašavaju da njihov rad nije namenjen da bude pravni savet, tako i autor ovog rada ističe da podaci navedeni u ovom radu ne mogu predstavljati pravni savet i da bi istraživači u slučaju pravnih nedoumica trebalo da se konsultuju sa pravnikom oko relevantnog zakonskog okvira i legalnosti konkretne vrste sakupljanja podataka.

U fazi paralelizacije, bilo je potrebno kreirati metodologiju serijalizacije HTML datoteka tako da njihov redosled bude identičan u svim jezicima. Inicijalna analiza datoteka pokazala je da se tekst nalazi u elementima *p*, *h1*, *h2*, *h3*, i *li*, sa izuzetkom oznaka segmenata poput „NAPOMENA“ ili „UPOZORENJE“, koje su se nalazile u elementima *div*, koje je bilo teže preuzeti s obzirom na to da većina elemenata *div* na sajtu ne sadrži tekst. Iz ovog razloga, odlučili smo da izuzmemo ove oznake segmenata iz preuzimanja.

S obzirom na to da su uputstva za isti model na različitim jezicima u najvećoj meri imala identičnu strukturu HTML elemenata, odabrali smo programsko generisanje TMX 1.4 datoteka pomoću *Python* skripti. Napominjemo da su u procesu izrade metodologije i pisanja kôda u našem istraživanju korišćeni modeli veštačke inteligencije (VI) kompanija OpenAI, Antropik i Gugl, posle čega je kôd ručno proveren, a hipoteze su istražene u relevantnoj literaturi. TMX datoteke, generisane pomoću modela VI, validirane su prema DTD šemi za TMX 1.4⁶ koju je izradila nekadašnja Asocijacija za lokalizaciju industrijskih standarda (engl.

⁶ <http://www.ttt.org/oscarStandards/tmx/tmx14.dtd>

Localization Industry Standard Association) (Savourel, 2002). Zaglavlje u prvobitnoj verziji TMX datoteka nije bilo u adekvatnom formatu, pa su sve datoteke ispravljene kako bi postale validne u odnosu na DTD šemu.

Jedna od najvećih prepreka u izradi korpusa Hilti-UM bila je validacija paralelizacije segmenata u TMX datotekama. S obzirom na veliki broj datoteka i na stepen poznavanja različitih jezika koji su sadržani u korpusu, ručna provera celog korpusa nije bila moguća. Iz ovog razloga, odlučili smo se za ručnu proveru celokupne srpsko-engleske paralelizacije i programsku validaciju ostalih jezičkih parova. Metodologija programske validacije je opisana u odeljku 3, a kao što je već pomenuto, jezici koji nisu zasnovani na ćirilicom i latiničnom pismu nisu mogli da budu uključeni u ovu validaciju, zbog čega je odlučeno da se izdvoje iz korpusa. Međutim, smatramo da opisana metodologija može biti korisna istraživačima koji se bave izdvojenim jezicima da sami izrade (i validiraju) slične paralelne korpuse uputstava i na tim jezicima.

6. Zaključak

S obzirom na potrebe istraživača u različitim poljima lingvistike i nastave jezika za kvalitetnim paralelnim tekstovima, naročito kada je u pitanju jezik struke, smatramo da metodologija opisana u ovom radu može biti korisna za izradu sličnih paralelnih korpusa. U odnosu na tradicionalne izvore tekstova, smatramo da veb-sajtovi koji prikazuju javno dostupne stranice sa identičnom strukturom na različitim jezicima predstavljaju znatno lakši izvor za izradu paralelnih korpusa, s obzirom na to da se tekst već nalazi u mašinski čitljivom formatu i da je struktura već uparena na nivou paragrafa. Smatramo da bi ovakvi resursi bili veoma značajni u oblasti prevođenja, nastavi prevođenja, različitim vrstama analize jezika, ali i kao vredni skupovi podataka u oblastima mašinskog učenja i veštačke inteligencije, pri čemu je potrebno razmotriti pravne i etičke aspekte automatskog sakupljanja građe u kontekstu pojedinačnog sajta, kao i ciljeve istraživanja i načine upotrebe resursa koji proishode iz istraživanja.

Bibliografija

- [1] Bañón, M., Chen, P., Haddow, B., Heafield, K., Hoang, H., Esplà-Gomis, M., Forcada, M. L., Kamran, A., Kirefu, F., Koehn, P., Ortiz Rojas, S., Pla Sempere, L., Ramírez-Sánchez, G., Sarrias, E., Strelec, M., Thompson, B., Waites, W., Wiggins, D., & Zaragoza, J. (2020). ParaCrawl: Web-Scale Acquisition of Parallel Corpora. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 4555–4567). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.417>
- [2] Krotov, V., Johnson, L., & Silva, L. (2020). Legality and Ethics of Web Scraping. *Communications of the Association for Information Systems*, 47, 539–563. <https://doi.org/10.17705/1CAIS.04724>
- [3] Lefer, M.-A. (2020). Parallel Corpora. In M. Paquot & S. Th. Gries (Eds.), *A Practical Handbook of Corpus Linguistics* (pp. 257–282). Springer International Publishing. https://doi.org/10.1007/978-3-030-46216-1_12
- [4] Lopresti, D. (2009). Optical character recognition errors and their effects on natural language processing. *International Journal on Document Analysis and Recognition (IJ DAR)*, 12(3), 141–151. <https://doi.org/10.1007/s10032-009-0094-8>
- [5] Marovac, U., Avdic, A., & Ljajić, A. (2021). Creating a stop word dictionary in Serbian. *Scientific Publications of the State University of Novi Pazar Series A: Applied Mathematics, Informatics and Mechanics*, 13, 17–25. <https://doi.org/10.5937/SPSUNP2101017M>
- [6] Norvig, P. (2009). Natural Language Corpus Data. In T. Segaran & J. Hammerbacher (Eds.), *Beautiful Data: The Stories Behind Elegant Data Solutions* (pp. 219–242). O'Reilly Media, Inc. <https://norvig.com/ngrams/ch14.pdf>
- [7] Savourel, Y. (Ed.). (2002). *TMX 1.4a Specification* [OSCAR Recommendation]. Localisation Industry Standards Association (LISA), OSCAR SIG. <https://www.ttt.org/oscarStandards/tmx/tmx14-20020710.htm>
- [8] Škorić, M., & Janković, N. (2024). New Textual Corpora for Serbian Language Modeling. *Infotheca*, 24(1). <https://arxiv.org/abs/2405.09250>
- [9] Unicode Consortium. (2025). *Unicode Standard Annex #29: Unicode Text Segmentation*. <https://www.unicode.org/reports/tr29/>
- [10] Varga, D., Halácsy, P., Kornai, A., Nagy, V., Németh, L., & Trón, V. (2007). Parallel corpora for medium density languages. In *Recent Advances in Natural Language Processing IV* (pp. 247–258). <https://doi.org/10.1075/cilt.292.32var>

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

Nikola Janković

Summary

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

Keywords: parallel corpora, machine translation, language tools, specialized language.

CIP - Каталогизација у публикацији
Народна библиотека Србије, Београд

811.163'322(082)(0.034.2)
004.8(082)(0.034.2)

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

Thematic collection of papers [Elektronski izvor] / Proceedings of the International Conference South Slavic Languages in the Digital Environment JuDig, Belgrade ; [organisers University of Belgrade - Faculty of Philology [and] Language Resources and Technologies Society JeRTeh] ; [editors Jasmina Moskovljević Popović, Ranka Stanković]. - Belgrade : University, Faculty of Philology, 2025 (Belgrade : University, Faculty of Philology). - 1 elektronski optički disk (CD-ROM) : tekst ; 12 cm

Sistemska zahtevi: Nisu navedeni. - Nasl. sa naslovnog ekrana. - Tiraž 10.

ISBN 978-86-6153-791-2

a) Јужнословенски језици -- Рачунарска лингвистика -- Зборници

COBISS.SR-ID 181147145

Edicija - naučni skupovi / Едиција – научни скупови / Edition – Scientific Conferences

www.fil.bg.ac.rs